
UNIVERSITY OF LIVERPOOL

# Comparative genomic analysis of *Neospora caninum* and *Toxoplasma gondii*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

**Eman Alshehri**

**January 2019**

## Acknowledgements

I would like to firstly thank my supervisors, Christiane and Neil, who both gave me the opportunity to do this project in the first place, and for their support and encouragement during the whole years. To Christiane, for your smile, understanding, patient, kindness, I will never forget you when you said; I am here to help, you will get it one day, thank you so much for believing in me. To Neil, thank you for your kindness and for the first cup of tea you made for me in my first day in university.

I'd also like to thank certain CGR members, thank you Dr Sam Haldenby, who helped me all the time, in particular for bioinformatic analysis during the last 12 months, thank you big boss. I'd also like to thank Dr David Starns for your support of me in our office during the last two years, for your smile and for listening to me as friend thank you Dave. I would also like to thank Amber, Ali, Balquees, Mara, Wazeera, Louise, Eva, Stalo, and all my friends in my home. Thanks, in particular to Elaine Daugan for being kind, support me to feel strong, optimistic and sorry for finishing your tissues.

Dad are you hearing me? I did it. It was hard after your leaving. As far as I know you, you are smiling now, I hope without tears. Thank you, Dad, for your support, love, for all the nice memories in my mind and my heart. Dad, my husband is amazing, brave, supportive, he helped me for doing this. Thank you my love for everything you did for me. For the best son in the world, one day you will read this, I did all of this for you, thank you for being nice all the time, for the everyday hugs, thank you for your advice; mum start with easy questions first. Thank you for your warm small hands. We did it.

For my family, my mum, all my family members, who without their support I would never finish this work. Finally, I want to thank me for believing in me, for doing the hard work, and for never stopping

# Comparative genomic analysis of *Neospora caninum* and *Toxoplasma gondii*

**Eman Alshehri**

## Abstract

*Toxoplasma gondii* and *Neospora caninum* are coccidian intracellular parasites that cause diverse pathological effects in humans and other warm-blooded vertebrates. Recent comparative genomic analyses of both species have identified genes specific to each of the species. These species-specific genes largely encoded proteins of unknown function. However, in other parasites such species-specific genes can be associated with host interaction and therefore we hypothesise that these genes play an important role in the parasite's ability to infect a wide range of vertebrate hosts and to avoid the host immune response. We revisited the comparative genomic analysis between *T. gondii* and *N. caninum* and identified 1,544 species-specific genes in *T. gondii* and 291 species-specific genes in *N. caninum*, extending on previously published data.

Furthermore, we used whole-genome sequencing to identify genetic variation (SNPs and CNVs) between six *T. gondii* and three *N. caninum* strains from different hosts and different global distributions. Our findings reveal that there were large expansions of multiple gene families that are known to be involved in pathogenesis, host range restriction, host-parasites interactions and disease severity in mice and likely also in humans. We further showed that there were additional members of the surface- antigen gene family (SRSs) genes that was significantly expanded in the *N. caninum* genome; this family is thought to mediate attachment to host cells and modulate the host immune response. This led us to support the hypothesis that the restriction in the host range of *N. caninum* might be associated with this gene family. Similarly, in *T. gondii*, a highly expanded gene family known as *Toxoplasma gondii* family proteins (TgFAMs), might be involved in adaptations, mechanism of immunity and transmission during sexual development in the definitive host.

We systematically analysed unmapped reads from whole genome sequencing resulting from regions missing, misassembled, or divergent from the two reference genomes. Using a *de novo* assembly pipeline on these unmappable reads, we identified novel genes in two *Neospora* and two *Toxoplasma* isolates, the majority of which were encoding proteins of unknown function. Our results provide valuable information about the differences between the genomes of *N. caninum* and *T. gondii*, which may underlie their divergence and will facilitate future approaches to expand the horizon of understanding the mechanism of virulence and defence strategies between the two closely related parasites.

## Table of contents

**Acknowledgements..... i**

**Abstract.....ii**

**Table of contents.....iv**

**List of figures.....ix**

**List of tables .....xi**

**Abbreviations.....xii**

**Chapter 1. Introduction**

**1.1 The parasites of the phylum *Apicomplexa*.....1**

**1.2 Phylogeny of *Toxoplasma gondii* and *Neospora caninum*.....2**

**1.3 The parasites *T. gondii and N. caninum*.....4**

**1.4 Impact of Toxoplasmosis and Neosporosis.....8**

**1.5 Life cycle of *T. gondii and N. caninum*.....9**

1.5.1 Tachyzoite stage of *T. gondii* and *N. caninum*.....10

1.5.2 Bradyzoites stage of *T. gondii* and *N. caninum*.....11

1.5.3 Oocysts stage in of *T. gondii* and *N. caninum*.....12

**1.6 Apical structure and secretory organelles of *Toxoplasma gondii* and *Neopsora caninum*.....12**

1.6.1 Rhoptries.....12

1.6.2 Micronemes.....14

1.6.3 Dense Granules.....14

**1.7 Host cell invasion by *T. gondii* and *N. caninum*.....15**

**1.8 Differences and similarities between *T. gondii* and *N. caninum*.....16**

1.8.1 Biological characteristics and ultrastructure differences.....16

1.8.2 Genome structure and genetic differences.....18

1.8.3 Previous comparative genomic and transcriptome analyses done on *T. gondii* and *N. caninum* parasites.....21

1.8.3.1 Surface antigens proteins.....24

1.8.3.2 Secreted encoded proteins.....25

**1.9 Strains variations of *T. gondii*.....26**

**1.10 Strains variations of *N. caninum*.....31**

**1.11 Genome sequencing.....33**

**1.12 Genome Assembly……..……..34**
**1.13 Single nucleotides polymorphisms detection and analysis…..... 35**
**1.14 Copy number of variation detection and analysis……..36**
**1.15 Aims of the thesis……..……37**

**Chapter 2. Materials and Methods…….……...39**
**2.1 Comparative genomic analysis of *T. gondii* and *N. caninum* parasites……..39**
2.1.1 Analysis of orthologous genes of *T. gondii* and *N. caninum* using OrthoMCL ……..39
**2.2 Strains selection……...41**
**2.3 Cell culture of the hosts and parasites……...44**
2.3.1 Vero cell passage……..……44
2.3.2 Parasite passage……..……..44
**2.4 Parasite purification……...45**
**2.5 Purification of total genomic DNA from tachyzoites……....47**
**2.6 Quality control assessment (QC)……...……48**
**2.7 Library preparation**……..**48**
**2.8 Whole genome sequencing (WGS)……51**
2.8.1 Read Processing and quality assessment of the raw sequence data……..…..51
2.8.2 Short reads alignment to the reference genome sequence……..……53
2.8.3 Manipulating the files with SAMtools and Picard Tools……..53
**2.9 Purified Mapped reads processing……...57**
2.9.1 SNPs discovery per sample……57
2.9.2 SNPs functional annotation…….……..59
**2.10 Gene ontology enrichment analyses……..60**
**2.11 Fold enrichment (REViGO)……..60**
**2.12 Purified unmapped reads processing……61**
2.12.1 Identification the composition of unmapped reads……61
2.12.2 *Do novo* assembly of unmapped reads pipeline……62
2.12.3 Assembly evaluation of the final outputs……..…….62
2.12.4 Integration of Genome Assemblies (Blob-tools)……...…62
**2.13 Gene finding and annotations pipeline……..……63**
**2.14 Copy Number of Variations estimation…….... 64**

**Chapter 3. Comparative genomic analysis of *T. gondii* and *N. caninum* parasites .69**

**3.1 Introduction**.....65

**3.2 Aims of this Chapter**.....66

**3.3 Results**.....67

**3.3.1 Identifying orthologues cluster between *T. gondii* and *N. caninum* parasites**.....67

**3.3.2 Identification of species-specific gene families in *T. gondii* and *N. caninum***.....73

3.3.2.1 Identification of SAG- related sequence (SRSs) genes.....73

3.3.2.2 Identification of Rhoptry (ROPs) genes.....75

3.3.2.3 Identification of Toxoplasma gene family (TgFAMs) genes.....78

3.3.2.4 Identification of Dense granule (GRAs) genes.....80

3.3.2.5 Identification of Micronemes (MICs) genes.....82

3.3.2.6 Identification of other gene families between species.....84

**3.4 Discussion**.....85

**Chapter 4. Use of multiple strain sequencing to define variants contributing to phenotyping changes among *N. caninum* isolates**.....87

**4.1 Introduction**.....87

**4.2 Aims of the chapter**.....90

**4.3 Results**.....91

**4.3.1 Data generation and sequence read alignment of the *N. caninum* isolates to the entire *N. caninum* Liverpool reference genome**.....91

**4.3.2 SNPs analyses**.....96

4.3.2.1 Comparison of SNPs rate in three strains of *N. caninum*.....96

4.3.2.2 Representation of the SNPs within the data.....101

**4.3.3** Investigation of frequency of genes that contain the SNPs within each strain.....103

**4.3.4** Investigation of genomic diversity within different sequence classes.....104

**4.3.5** High impact SNPs unique to specific strain.....108

**4.3.6** Investigating of most diverged genes within each strain.....111

**4.3.7 GO enrichment analysis**.....120

4.3.7.1 Pathways enriched in genes containing predicted modifying impact SNPs……...120
4.3.7.2 Pathways enriched in genes containing predicted moderate impact SNPs……...……122
4.3.7.3 Pathways enriched in genes containing predicted high impact SNPs……..……..123
**4.3.8 Copy number of variations and the gene annotations of *Neospora caninum* strains**……127
**4.3.9 *De novo* assembly analysis of unmapped reads ……**.....131
**4.3.10 Gene finding and annotation**……...……..137
**4.4 Discussion**……...……142

**Chapter 5. Use of multiple strain sequencing to define variants contributing to phenotyping changes among *T. gondii* isolates**……148
**5.1 Introduction**……148
**5.2 Aims of the chapter**……149
**5.3 Results**……..151
**5.3.1 Data generation and sequence read alignment of the *T. gondii* isolates to the entire *T. gondii* ME49 reference genome……..…**151
**5.3.2 SNPs analyses**……..156
5.3.2.1 Identification of SNPs between the six lineages of *T. gondii* ……...…156
5.3.2.2 Distribution and density of SNPs in *T. gondii* isolates……...……161
**5.3.3** Investigation genomic diversity within different sequence classes .163
**5.3.4** Investigation of frequency of SNPs in the most diverse genes within each strain……173
**5.3.5 GO terms analysis**.……..……179
5.3.5.1 Pathways enriched in genes containing predicted modifier impact SNPs……179
5.3.5.2 Pathways enriched in genes containing predicted moderate impact SNPs……...…..182
5.3.5.3 Pathways enriched in genes containing predicted high impact SNPs...184
**5.3.6 Investigating high impact SNPs unique to each strain of *T. gondii*……187**

5.3.6.1 High impact SNPs identification in key biologically relevant Gene families in distinct *T. gondii* strains…….....191
5.3.6.1.1 Surface antigen gene family (SAG)…….191
5.3.6.1.2 Rhoptry kinase proteins (ROPs)……...194
5.3.6.1.3 Dense granule (GRAs) genes…….……..197
5.3.6.1.4 Micronemes (MICs) genes……...198
5.3.6.1.5 Toxoplasma gene family (TgFAMs) genes……..198
5.3.6.1.6 Lysine- Arginine rich Unidentified Function family (KRUFs)……..199

**5.3.7 Copy number of variations and the gene annotations of *T. gondii strains***……....204
5.3.7.1 GO terms analysis of the genes with CNV in *T. gondii* strains……213
**5.3.8 *De novo* assembly analysis of unmapped reads**……...217
**5.4 Discussion**…….226

**Chapter 6.**
**6.1 General discussion**……...**224**
**6.2 Concluding remarks and contributions to the field**…… **231**
**6.3 Future works and scientific contributions of achievements.232**

**References**…….**237**

**Appendices**……....**253**

# List of figures

**Chapter 1:**

**Figure 1.1……10**

**Chapter 2:**

**Figure 2.1……44**
**Figure 2.2……47**
**Figure 2.3……50**
**Figure 2.4……55**
**Figure 2.5……58**
**Figure 2.6……59**
**Figure 2.7……60**
**Figure 2.8……62**

**Chapter 3:**

**Figure 3.1……71**
**Figure 3.2……72**
**Figure 3.3……72**

**Chapter 4:**

**Figure 4.1……95**
**Figure 4.2……99**
**Figure 4.3……100**
**Figure 4.4……102**
**Figure 4.5……105**
**Figure 4.6……106**
**Figure 4.7……107**
**Figure 4.8……113**
**Figure 4.9……114**
**Figure 4.10……115**
**Figure 4.11……116**
**Figure 4.12……117**
**Figure 4.13……118**
**Figure 4.14……119**
**Figure 4.15……124**
**Figure 4.16……125**
**Figure 4.17……126**
**Figure 4.18……129**
**Figure 4.19……130**
**Figure 4.20……134**
**Figure 4.21……135**
**Figure 4.22……136**
**Figure 4.23……138**

**Chapter 5:**

**Figure 5.1……..155**
**Figure 5.2……..158**
**Figure 5.3……..160**
**Figure 5.4……..165**
**Figure 5.5……..166**
**Figure 5.6……..167**
**Figure 5.7……..168**
**Figure 5.8……..169**
**Figure 5.9……..170**
**Figure 5.10……171**
**Figure 5.11……172**
**Figure 5.12……177**
**Figure 5.13……....181**
**Figure 5.14……....183**
**Figure 5.15……....185**
**Figure 5.16……....186**
**Figure 5.17……....190**
**Figure 5.18……....201**
**Figure5.19…….…….... 202**
**Figure 5.20……...……..203**
**Figure 5.21……207**
**Figure 5.22……208**
**Figure 5.23……209**
**Figure 5.24……210**
**Figure 5.25……211**
**Figure 5.26……212**
**Figure 5.27……214**
**Figure 5.28……215**
**Figure 5.29……216**
**Figure 5.30……220**
**Figure 5.31……221**
**Figure 5.32……222**
**Figure 5.33……223**
**Figure 5.34……224**

# List of tables

**Chapter1:**

**Table 1.1……...7**
**Table 1.2……...17**
**Table 1.3…...……..19**
**Table 1.4……...20**

**Chapter 2:**

**Table 2.1……46**

**Chapter 3:**

**Table 3.1……70**
**Table 3.2……74**
**Table 3.3……77**
**Table 3.4……79**
**Table 3.5……81**
**Table 3.6……83**

**Chapter 4:**

**Table 4.1……93**
**Table 4.2……94**
**Table 4.3……...97**
**Table 4.4……....98**
**Table 4.5……...110**
**Table 4.6……..…133**
**Table 4:7……..139**
**Table 4.8……..140**

**Chapter 5:**

**Table 5.1……...153**
**Table 5.2……...154**
**Table 5.3……...157**
**Table 5.4……...159**
**Table 5.5……...178**
**Table 5.6……...189**
**Table 5.7……...200**
**Table 5.8……...219**
**Table 5.9……...225**

## Abbreviations

**BLAST** basic local alignment search tool

**BWA** Burrows Wheeler Aligner

**bp** base pairs

**E-value** Exponent value in BLAST search

**GO** Gene ontology

**GPI** glycophosphatidyl inositol

**GRA** dense granule protein

**IGV** Integrative Genomics Viewer

**kb** kilo base

**Mbp** mega base pair

**MIC** Microneme protein

**MLST** multi-locus sequencing typing microsatellites

**MS** microsatellites

**mRNA** Messenger RNA

**NGS** Next Generation Sequencing

**PBS** phosphate buffered saline

**PVM** parasitophorous vacuole membrane

**RFLP** restriction fragment length polymorphism

**RIN** Integrity number

**RON** rhoptry neck protein

**ROP** rhoptry protein

**QC** quality control

**SAG** surface antigen

**SAM** Sequence Alignment Map

**SNP** single nucleotide polymorphism

**SRS** SAG1-related sequence

**TMRCA** Time to Most Recent Common Ancestor

**VCF** Variant Call Format

# Chapter 1: Introduction

## 1.1 The parasites of the phylum *Apicomplexa*

The *Apicomplexa* is a protozoan phylum comprising around 6000 named species and it might that be some species are still undiscovered (Seeber and Steinfelder, 2016). There are three major taxonomic classes including coccidia, haemosporidians, piroplasmids and gregarines that belonging to this phylum. All members of the Apicomplexa share a common feature of an apical complex in one or more stages of the life cycle (Morrison, 2009). Globally, many epidemiologic studies have shown that the vast majority of the *Apicomplexan* parasites are pathogenic, causing serious infection diseases such as malaria caused by *Plasmodium* species that is considered a life threatening disease that kills a large number of people every year. *Plasmodium falciparum* usually has the highest level of severity compared to other species (Pain *et al.*, 2008; Wellems, Hayton and Fairhurst, 2009; Ashley, Pyae Phyo and Woodrow, 2018). Another is toxoplasmosis, caused by *Toxoplasma gondii* that can infect a wide range of hosts including humans and is a particular danger during pregnancy and in immunocompromised patients and animals (Tenter, Heckeroth and Weiss, 2000; Dubey and Jones, 2008; Robert-Gangneux and Dardé, 2012).

Neosporosis is one of the pathogenic diseases caused by an apicomplexan parasite; *Neospora caninum* infects a wide range of livestock causing abortion in cattle and it can be also have an economic impact on the dairy industries such as milk production (Dubey, Schares and Ortega-Mora, 2007a; Donahoe *et al.*, 2015). In this regard, it has been found that there are other pathogens belonging to this phylum, such as *Cryptosporidium*, *Eimeria*, *Babesia, Theileria* and *Isospora,* that have different patterns of morbidity, mortality, morphology, host ranges, transmission strategies, invasion mechanism ,medical, agriculture and economic importance (Morrissette and Sibley, 2002; Francia and Striepen, 2014).

## 1.2 Phylogeny of *Toxoplasma gondii* and *Neospora caninum*

Recent evolutionary estimation was performed to provide higher resolution of the Apicomplexa parasites to determine their taxonomy and phylogeny framework, more specifically between the two closely related coccidian parasites; *T. gondii* and *N. caninum* that have been dramatically changed over time. It has been previously estimated that their divergence time was between 12 and 80 million years ago (Mya) (Su *et al.*, 2003; Ricklefs and Outlaw, 2010). Phylogenetic evidence provided by Paibomesai *et al*., (2010) indicated that there were groups of specific genes called clock-like genes that play a key role in the life history of the organisms due to frequent accumulation of high proportions of Single Nucleotide Polymorphisms (SNPs) distributed among the genomes, which result in different phenotypes (Paibomesai *et al.*, 2010). Recent divergence time comparison has been done by Reid *et al*., (2012) and revealed that the two parasites and *Plasmodium* species differed in divergence time due to the specification of both parasites and their definitive hosts around 28 million years ago (Kumar, 2005; Reid *et al.*, 2012).

Limited information was available about the relationships between different phylogenetic trees and the absence of fossil records for species especially after introducing high quality advance sequencing technologies that improve the overall genome resequencing for both species. One of the recent bioinformatic tools for assessing the divergence times and clock like speciation between different species is called Time Tree (http://www.timetree.org), which is a databases that can be used to determine the phylogenetic relationships and divarication degree between organisms assembled from several published molecular studies. (Hedges, Dudley and Kumar, 2006; Hedges *et al.*, 2015).

Recent calculating using the Time Tree tool confirmed that the speciation of the parasites occurred after the divergence their definitive hosts around 54 -67 Mya (Reid *et al.*, 2012). The allelic frequencies can potentially alter the fitness of the parasites and introduce new isolates with new biological traits over a long period of time (Sibley *et al.*, 2002; Barragan and Sibley, 2003; Li *et al.*, 2003; Nath and Sinai, 2003; Boothroyd, 2009; Raz and Tannenbaum, 2010). An additional recent phylogenetic tree of *T. gondii* and *N. caninum* species used more samples of distinct isolates for both

parasites and generate different phylogenetic trees over time by using different methods: Microsatellite genotyping and genome sequencing was used to determine the genetic diversity not only between the two parasites but also within the isolates of each species (Regidor-Cerrillo *et al.*, 2013; Lorenzi *et al.*, 2016) providing further support for the hypothesis that the genetic exchange with varied clusters was directly depend on host migration events and parasite recombination events in *T. gondii* and *N. caninum.* The recent study by Calarco (2018) offers a further comprehensive analysis of SNPs in the genes that play a significant role in virulence between two strains of *N. caninum* (Calarco, Barratt and Ellis, 2018). Geographically, there was a significant association between the population structure of *T. gondii* with the geographical segregation as we will describe later in section 1.9, However, there was no direct link between the *N. caninum* isolates and geographical distribution as we will explain in depth in section 1.9 section. The validity of the theory that molecular clock and evolution was changed in the species, influenced by factors such as number of variations, expansion of multiple gene families, genome arrangements has advanced due to advanced next generation sequencing methods (NGS) for sequencing and resequencing genomes. Theses show the forces shaped the population structures in species and improved the evolutionary timescales of *T. gondii* and *N. caninum* genomes (Paterson, Vogwill, Buckling, Benmayor, Andrew J Spiers, *et al.*, 2010; Shwab *et al.*, 2014; Auld and Tinsley, 2015; Donahoe *et al.*, 2015; Thankaswamy-Kosalai, Sen and Nookaew, 2017).

### 1.3 The parasites *Toxoplasma gondii* and *Neospora caninum*

*Toxoplasma gondii* and *Neospora caninum* are two closely related protozoan parasites that are distributed worldwide. Both organisms are members of the phylum Apicomplexa, class Sporozoa and subclass Coccidia, which can infect a wide range of hosts as mention earlier in section 1.1. It was only in 1984 that *N. caninum* was distinguished from *T. gondii* as a separate species and described and characterised by Dubey and colleagues (Dubey *et al.*, 1988; Dubey, Lindsay and Speer, 1998).In addition, *T. gondii* is the only known species in the genus *Toxoplasma*; however, there are two known species with the genus of *Neospora; N. caninum* and *N. hughesi*. The two species have the same morphological structure, but there are some differences in ultrastructure, genome structure, host preference, pathogenies and transmission between these two species(Schares *et al.*, 1998; Hill and Dubey, 2002; Dubey, Schares and Ortega-Mora, 2007a; Hassan *et al.*, 2012; Reid *et al.*, 2012; Regidor-Cerrillo *et al.*, 2013).

There are three major distinct clonal lineages of *T. gondii* with varied biological characteristics includes growth rate, virulence, migration, mechanism of invasion and geographical distribution named type I, II and III(Howe and Sibley, 1995a; Lehmann *et al.*, 2000; Saeij, Boyle and Boothroyd, 2005).Recently, it has been found that there was a fourth clonal type known as type 12 (Khan *et al.*, 2006a; Dubey *et al.*, 2011; Khan, Dubey, *et al.*, 2011; Su *et al.*, 2012). It has been reported that there is a high level of similarity between the strains of *T. gondii,* although isolates show biological and genetic variations that contribute to significant variability in virulence. One of the most considerable phenotypic differences is the pathogenicity that is usually determined based on the intraperitoneal inoculation of tachyzoites (infective stage) in experimental mouse models and which defines virulence. Type I isolates such as GT1 and RH strains were highly virulent having 100% probability of causing death ($LD_{100}$) in the hosts and in the laboratory mouse models. The other two types, (II and III) are considered non-virulent, for instance ME49 and VEG strains which have a less lethal dose ($LD_{50}$) in mice. Data collected from human cases of toxoplasmosis show a high incidence of severe acquired infection attributed to type I strains, which are acutely virulent with fast rates of replication *in vitro* and reach higher tissue burden in muscle and brain tissue (Saeij, Boyle and Boothroyd, 2005; Khan *et al.*, 2006a; Behnke *et al.*,

2011; Dubey *et al.*, 2011; Khan, Dubey, *et al.*, 2011; Su *et al.*, 2012) (Table 1.1). Indeed, there were noticeable variations even between the strains with same genotypes. In type II and III strains, many experimental studies have proved they were considered less virulent than type I. A study conducted in the USA showed that most cases of congenital toxoplasmosis were associated with type II (Sibley and Boothroyd, 1992). In agreement with this finding, genotyping study of congenital samples in France concluded that almost all the cases belonged to type II (Ajzenberg *et al.*, 2004). Generally, type III, represented by the VEG strain was considered less virulent and was significantly associated with animal hosts. However, a new systematic analysis of existing genotypes of *T. gondii* strains that were isolated from animals revealed that additional strains showed a high level of virulence even though they belonged to the type III, such as strain P89 (Behnke *et al.*, 2014).

Several experimental studies of crosses between virulent isolates (I x III, II x II, IIxIIII and I x II) have shown that there was genetic exchange in recombinant progeny (Saeij, Boyle and Boothroyd, 2005). Analyses of these progeny revealed the allelic variations between the lineages. These findings confirmed that the virulence trait is heritable and multigenic. It also appears that there was a significant link between the genetic background of specific strains with loci located on particular chromosomes and acute virulence among the parental types (Dardé, Ajzenberg and Su, 2013). In addition, genome-scale analyses and gene knockout of distinct isolates identified specific multiple genes that conferred enhanced pathogenicity among clonal parental and recombinant strains. In comparison to other features such as migration, serum response and growth rate have shown a variety of significant degree of variation that have been statistically associated with chromosomes, VIIa, XI and XII by genome-wide mapping (Sibley *et al.*, 2002; Khan *et al.*, 2005; Bontell *et al.*, 2009; Su *et al.*, 2012; English, Adomako-Ankomah and Boyle, 2015). In this regard, the phenotypic traits are not only controlled by the single locus, but there are multiple loci that can modulate the traits. This is seen on strain-specific chromosomes that cause a dramatic difference in the degree of virulence between types. Consistent with this, each strain has a specific complement of secreted effectors proteins that play a role in host-parasite interaction and in the pathways that affect the virulence. Several investigations on experimental strains have been performed to identify the virulence genes to determine the strain - specific virulence phenotypes, life cycle and host ranges in a wide range of habitats

(Daniel K Howe *et al.*, 1997; Fuentes *et al.*, 2001). Previous comparative studies pointed to strain - specific differences in other phenotypic traits that caused changes in the virulence profiles in different strains of *T. gondii*, which focused on the acute virulence between clonal linages and assessed growth rate, migration, gene expression and other combination of effectors that may influence different pathways of host-parasites interaction (Dardé, 2004; Laliberté and Carruthers, 2008). There was significant evidence that the growth rate differed between strains. Specifically, the growth rate of type I strains has been shown to be dramatically faster than for other types in mouse models (Radke *et al.*, 2001). The discrepancy in growth rate between the parental lines might be explained by understanding the host immune mechanism that is modulated by *T. gondii* strains (Sibley *et al.*, 2002; Saeij, Boyle and Boothroyd, 2005).

Assessing Quantitative Trait Loci (QTL) of the progeny from a I x III genetic cross *in vitro*, showed a significant peak on chromosomes VIIa, XI, XIII and Ia. Furthermore, migration differed significantly among in the parental strains and progeny with higher rate of potential migration in type I than other types based on QTL mapping evidences that mapped on the chromosome VIIa (Taylor *et al.*, 2006). In addition, there were multiple proteins and other novel secreted virulence factors that showed dramatic changes in the mechanisms of infection and the level of expression influenced virulence and modulated the host immune pathways in different clonal types burden in a wide range of hosts to ensure strains transmission as we will explain further in section 1.7 (M. E. Grigg *et al.*, 2001; Dubremetz and Lebrun, 2012). Importantly, it is useful to know how the secreted kinases are regulated during the host – pathogen interaction and to find determents that are directly involved in host cell mechanisms to identify strain-specific virulence differences in different hosts. Clearly then, these results suggest that each of the clonal isolates of *T. gondii* have unique characteristics of virulence loci and a key different complement of specific alleles that clearly demonstrated both strain specificity and host range, which had a significant impact on the geographic separation of diversity of *T. gondii* strains in different resident regions in animals and human (Su *et al.*, 2012; Wang *et al.*, 2012; Shwab *et al.*, 2014).

**Table 1.1:** Summary of biological and epidemiological characteristics of the three main clonal lineages of *Toxoplasma gondii* (Saeij, Boyle and Boothroyd, 2005).

| Type | Description |
|---|---|
| **Type I (***GT1-RH***)** | **Highly virulent** for mice: death of all mice inoculated with less than 10 tachyzoites. *In vitro:* high rate of multiplication, reduced interconversion tachyzoite-bradyzoite. |
| **Type II *(ME49)*** | **Medium-virulent** for mice: chronic infection with persistence of tissue cysts. *In vitro:* slow rate of multiplication, easier interconversion tachyzoite-bradyzoite and formation of cysts. |
| **Type III** *(VEG)* | **Avirulent.** Rare among. *T. gondii* isolates originating from Europe and USA. More frequent among isolates originating from wild animals. |

## 1.4 Impact of Toxoplasmosis and Neosporosis diseases

Toxoplasma is one of the most successful zoonotic parasites, due to its efficient transmission. It causes a variety of disease in humans, as well as neonatal mortality and spontaneous abortion in livestock worldwide (Tenter, Heckeroth and Weiss, 2000). The most dramatic clinical manifestations in humans and animals are abortion or foetal damage depending on the pregnancy stage. Sever symptoms that have been observed are hydrocephalus, congenital toxoplasmosis, neurological deficiencies, deafness, seizures, retinochoroiditis that lead to eye lesions and blindness. *T. gondii* is life threating in AIDS patients (Tenter, Heckeroth and Weiss, 2000; Jones *et al.*, 2001; Robert-Gangneux and Dardé, 2012; Flegr *et al.*, 2014). In animal species, recent estimates provide evidence of the high prevalence of toxoplasmosis and differed between different countries (Flegr *et al.*, 2014). However, the prevalence of *T. gondii* was associated with genetic diversity, age, environmental factors and cultural and eating habitats (Dubey, Lindsay and Speer, 1998; Ajzenberg *et al.*, 2004).

Recent outbreaks of toxoplasmosis in sheep demonstrate high prevalence of infection with *T. gondii*, during early pregnancy (Buxton *et al.*, 1989; Lindsay and Dubey, 1989). There are various serological methods designed to detect the level of antibodies in individual sheep that were infected with Toxoplasmosis (Pappas, Roussos and Falagas, 2009).

*N. caninum* is the causative agent of Neosporosis; a disease that mainly infects cattle and dogs which has a significant economic impact in livestock industries. Many experimental studies confirm that there is a strong association between abortion in cattle and infection with *N. caninum* (Dubey, Schares and Ortega-Mora, 2007a).

### 1.5 Life cycle of *T. gondii* and *N. caninum*

Both *T. gondii* and *N. caninum* have similar life cycles. They have complex lifecycles, which include an asexual stage in the intermediate hosts and sexual stages in the definitive hosts. There are two modes of transmission in *N. caninum*; vertical transmission where it is transmitted from mother to foetus via the placenta and horizontal transmission via oocysts and tissue cysts. In contrast, in *T. gondii* the only mode for transmission is the horizontal mode (Dubey, Schares and Ortega-Mora, 2007b). There are three infectious forms of *T. gondii*, namely the tachyzoites (individually and in groups), bradyzoites (in tissue cysts), and oocysts (Tenter, Heckeroth and Weiss, 2000). The definitive host, which is the cat in *T. gondii* and dog in *N. caninum*, is infected through digestion of contaminated food or water, or from transplacental transmission of tachyzoites to developing fetuses (Dubey, 2013). The parasites are then infected epithelial cells of the definitive host's small intestine then starting the sexual cycle development in the gut epithelium of cat or dogs, producing millions of oocysts (Dubey *et al.*, 1988; Barber and Trees, 1998).

When the intermediate hosts ingest contaminated substances with sporulated oocysts, sporozoites are released from the oocysts and differentiate into tachyzoites, which spread to tissues and organs via blood or lymph to start the extra intestinal stage of the parasites (Montoya and Liesenfeld, 2004; Hide *et al.*, 2009; Sullivan and Jeffers, 2012). Tachyzoites multiply in parasitophorous vacuoles (PV) asexually in intermediate hosts through a process called endodyogeny. Cats shed oocysts from 4 to 11 days' post infection that survive for long periods due to their high resistance against adverse environmental conditions (Frenkel, Dubey and Miller, 1970), which results in widespread contamination of with the infective stage environments. In the latent stage (chronic form) of infection, tachyzoites are converted into the cyst forming bradyzoite stages, which are mostly found in the nervous system and muscles (Nath and Sinai, 2003).

### 1.5.1 Tachyzoite stage of *T. gondii* and *N. caninum*

The tachyzoite is crescent-shaped, approximately 2 × 7 µm with a pointed anterior end, which defines the direction of motility. It contains several organelles such as (subpellicular microtubules, conoid, inner membrane complex), secretary organelles (rhoptries, micronemes, dense granules), mitochondrion, apicoplast, nucleus, endoplasmic reticulum, Golgi apparatus and ribosomes, all surrounded by a complex membrane structure called the pellicle (Robert-Gangneux and Dardé, 2012). The parasite enters the host cell by active penetration of the cell membrane and becomes surrounded by a parasitophorous vacuole (PV) that protects it from host defense mechanisms. Tachyzoites convert to a slow growing form called bradyzoites, which replicate every 6 to 8 hours within host cells in the parasitophorous vacuole (PV) (Figure 1.1).

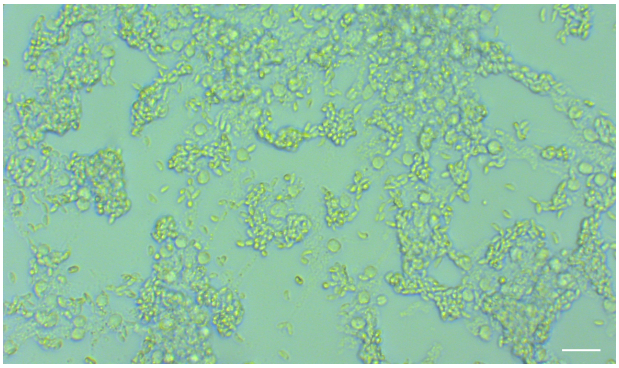


**Figure 1.1:** *T. gondii* tachyzoites (crescent-shaped) emerge from host cell. The image was taken from our experiment during tissue culture passaging of *T. GT1* strain, the tachyzoites were clustered after releasing from the host cell. Scale bar =10 µm.

### 1.5.2 Bradyzoites stage of *T. gondii* and *N. caninum*

Bradyzoites are an intracellular, slow replicating, quiescent form of the parasites and are an indication of the chronic stage of the disease. The crescent-shaped bradyzoites are 5-8.5 × 1-3 μm in size. Structurally, there are some differences between bradyzoites and tachyzoites. Bradyzoites are less slender than tachyzoites, the nucleus is located more towards the posterior end of the parasite compared to the centrally located tachyzoites nucleus. Bradyzoites contain several amylopectin granules and replication does not cause the rupture of host cells (Dubey, Lindsay and Speer, 1998; Lindsay *et al.*, 2006). Tissue cysts of *T. gondii* are usually spherical shaped. The size of bradyzoites can reach up to 100 μm containing hundreds of parasites. In contrast, the tissue cysts in *N. caninum* are rounded or oval shaped and can reach up to 107 μm in size and the thickness of the cyst wall can reach up to 4 μm and can persist during infection in the intermediate host without causing significant clinical manifestation(Dubey, 2005; Kim and Boothroyd, 2005)

### 1.5.3 Oocysts stage in of *T. gondii* and *N. caninum*

Oocysts shed from naturally infected definitive hosts as unsporulated form and then form two sporocysts, containing four sporozoites(Dubey, Lindsay and Speer, 1998). Dogs and cats shed oocysts after ingestion of tissue contaminated from the intermediate hosts or by ingestion of infectious oocysts from the environment. It has been reported that one cat shed millions of oocysts in *T. gondii* (Dubey, Lindsay and Speer, 1998; Hill and Dubey, 2002; Dubey, Schares and Ortega-Mora, 2007a). Gondim *et al.* 2005 provided evidence that the number of oocysts shed differs between *N. caninum* and *T. gondii*. In *N. caninum*, dogs have the ability to shed approximately 500,000 oocysts. The comparatively low frequency of oocyst sheds from dogs could be related to several factors such as, the age of dogs and the immunity of the host (Schares *et al.*, 1998; Gondim *et al.*, 2004). It is epidemiologically important to know the risk factors that increase the prevalence of infection. Sporulation occurs outside the definitive hosts and environmental conditions play a significant role in oocyst survival. The sporulation is considerably different among countries, regions and cities. In recent studies, it has been observed that moist and warm conditions might increase the survival of the oocyst (Frenkel, Dubey and Miller, 1970; Dubey, 1998, 2006).

## 1.6 Apical structure and secretory organelles of *T. gondii* and *N. caninum*

All members of the phylum of *Apicomplexa* have a crescent shaped structure called the apical complex in one or more stages of their life cycles, which plays a significant role in the host cell invasion process (Dubey *et al.*, 1998). In both parasites, there are a secretory organelles that include; Rhoptries (ROPs), Micronemes (MICs) and Dense granules (GRAs) (Carruthers and Boothroyd, 2007).

### 1.6.1 Rhoptries (ROPs)

The rhoptries are unique secretory organelles that are located in the anterior pole of the tachyzoites of *T. gondii* and *N. caninum* (L. David Sibley *et al.*, 2009; Kemp, Yamamoto and Soldati-Favre, 2013; Jensen *et al.*, 2015). In most species studied to date, the rhoptries proteins play a key role in host cell invasion, related to sequentially release of the parasite cell contents during invasion.

Rhoptry proteins separate into two intra-organellar compartments; the neck and the bulb. A detailed proteomic analysis of purified rhoptry proteins revealed thirty-eight novel rhoptry proteins. In addition, the rhoptry proteins could separate between bulb (ROP) and neck (RON) locations (Bradley *et al.*, 2005; L. David Sibley *et al.*, 2009). The rhoptry neck proteins (RONs) are located in the neck of the rhoptry organelle, conserved among *Apicomplexan* species and are involved in the formation of the moving junction (MJ) by forming a stable complex involving rhoptry neck proteins (TgRON2, TgRON4, TgRON5 and TgRON8) and apical membrane antigen 1 (TgAMA-1) (Camejo *et al.*, 2014). The ROPs proteins are involved in developing the dynamic compartment called parasitophorous vacuole membrane (PVM), which is involved in parasite survival. Importantly in this context, the ROP proteins are injected into the host cytoplasm but released later than the RON proteins in the invasion process (Sibley, 2011; Clough and Frickel, 2017). More specifically, ROP5, ROP16 and ROP18 are involved in virulence and modulating the host immune mechanism in *T. gondii* and its close relatives. In previous studies, it has been reported that ROP18 was a key virulence effector which can, by phosphorylating and inactivating phosphorylate immunity-related GTPases (IRGs) actively increase killing within PVM (Hunter and Sibley, 2012; Lei *et al.*, 2014; Jensen *et al.*, 2015).

A comparison of the three major clonal lineages of *T. gondii* revealed that there were strain differences in virulence (Saeij, Boyle and Boothroyd, 2005; El Hajj *et al.*, 2007; Behnke *et al.*, 2015). A high level of mortality was noticed in type I and II but, in type III no expression of ROP18 was noticed in a mouse model (Niedelman *et al.*, 2012; Behnke *et al.*, 2015; Shwab *et al.*, 2016). It has been found that ROP18 was code for by a pseudogene in *N. caninum* caused by several stop codons in the DNA sequence and thus might contribute to the differences between *T. gondii* and *N. caninum* (Reid *et al.*, 2012). In addition to ROP18, ROP5 is a pseudokinase that exists in three diverged isoforms (A-B-C) present in a different number of copies in *T. gondii* and *N. caninum* parasites. With the recent sequencing of both *T. gondii* and *N. caninum* genomes, it has been revealed that there is a significant role of ROP5 in virulence and host-parasite interaction through regulation of IRG binding working with ROP18 and with other members of the rhoptry kinase family (ROPKs). ROP16 is an active kinase that has direct association with the host-immune responses by altering the host cell transcription.

This differs between *T. gondii* strains (Ong, Reese and Boothroyd, 2010). It has been reported that the role of ROP16 is in controlling host signalling by activating signal transducer activator transcription (STAT) and activator of transcription (STAT3 and STAT6) during infection (Boyle *et al.*, 2006; Butcher *et al.*, 2011).

### 1.6.2 Micronemes

Micronemes are the smallest organelles that are found in Apicomplexan parasites that secrete their contents into the host cell stimulated by the mobilization of the parasites calcium. The microneme proteins are mostly known as MICs. Early investigations demonstrated the involvement of micronemes in host-cell invasion, binding and motility in *T. gondii* and *N. caninum* (Dubremetz and Lebrun, 2012). Recent analyses have contributed towards better understanding of the functional domains on those proteins. Almost all MICs have at least one adhesive domain that allows attachment to the target cell types in different life stages, which can bind TgMIC1, TgMIC4 and TgMIC6 to build a stable complex that is important to host cell invasion (Carruthers, 2002; Tonkin *et al.*, 2010; Sidik *et al.*, 2016)

### 1.6.3 Dense Granules

The dense granules (GRA) are spherical electron dense vesicles which vary in number depending on the life cycle stage of the parasites. GRA proteins, which are important to contributing and maintaining the PV, located in the cyst wall to be involved in different interaction between the parasites and the host cell. A recent study reported that GRA proteins are grouped into two types: Firstly, Conical GRA proteins such as GRA1-12, GRA14, GRA 20-23 and GRA 25. The second are the recently annotated GRA-like proteins such as GRA11, and GRA12 (Cesbron-Delauw, 1994; Ferguson *et al.*, 1999; Michelin *et al.*, 2009; Gold *et al.*, 2015).

### 1.7 Host cell invasion by *T. gondii* and *N. caninum* parasites

More recent studies have revealed that Micronemes (MICs), Rhoptries (ROPs), and Dense granules (GRAs) secrete their contents in a precise sequence of events during host cell invasion. To begin with, the invasive stage (tachyzoite) invades the target host cell. The tachyzoite moves to the host cell surface by gliding motility machinery that involve the actin-myosin system and is critical for successful penetration into the host plasma membrane (Carruthers and Sibley, 1997; Meissner., Reiss., Viebig., Carruthers., Tomavo. and and Soldat., 2002; Kim and Weiss, 2004; Carruthers and Boothroyd, 2007). Other group of genes that widely participate in the host cell invasion, mediate attachment, immune pathology and regulate the parasite's virulence of infection are the SRSs genes (SAG1-related sequences) found in the cell surface of both parasites (Lekutis *et al.*, 2001; Jung, Lee and Grigg, 2004; Risco-Castillo *et al.*, 2011; Wasmuth *et al.*, 2012). Further sets of proteins are secreted from micronemes, rhoptries and dense granules, which are needed to perform their individual tasks during the host cell invasion process. Firstly, the MICs proteins are secreted to start the first attachment step between the parasites and the host cell surface and penetration into the host cell. Following that, the ROPs proteins are released into the PV. Finally, the dense granules proteins are secreted into the lumen to modify the PV for acquisition of nutrients from the target cell (Kemp, Yamamoto and Soldati-Favre, 2013; Talevich and Kannan, 2013a).

A key stage in host cell invasion is the formation of a tight molecular structure between parasite and host cell membranes known as the moving junction (MJ). More specifically, four distinct secreted proteins released from different secretory organelles, (RON2, RON4, RON5, RON8) that along with AMA1 forme the moving junction complex (MJ), which contributes to host cell invasion through forming the PVM (Carruthers and Boothroyd, 2007; Tyler, Treeck and Boothroyd, 2011; Takemae *et al.*, 2013).

## 1.8 Differences and similarities between *T. gondii* and *N. caninum*

### 1.8.1 Biological and ultrastructure differences

*T. gondii* and *N. caninum* are closely related phylogenetically, have similar structures, share many common morphological, genetic, biological features and both infect a wide range of vertebrate hosts (Hill and Dubey, 2002; Al-Qassab, Reichel and Ellis, 2010). Despite these similarities, the two species differ in their definitive host: the definitive host of *T. gondii* is the cat (Felidae family) and in *N. caninum* is the dog (Canidae family). *T. gondii* can infect essentially any warm-blooded host and up to one third of humans are chronically infected with *T. gondii*. However, *N. caninum* is more restricted in host distribution when compared to *T. gondii.* It can infect cattle and dogs and may potentially infect horse, chickens, red foxes, sparrows, goats, sheep, and white –tailed deer (Salehi, Gottstein and Haddadzadeh, 2015). Strikingly, *T. gondii* can infect humans but *N. caninum* does not. This may be due to many factors such as, the number of genes and the role of secreted virulence proteins (Reid *et al.*, 2012; DeBarry and Kissinger, 2014; Reid, 2015; Lorenzi *et al.*, 2016). The summary of the main differences between the two closely related species, *T. gondii* and *N. caninum* are summarized in Table 1.2.

**Table1.2**: Summary of main biological differences between *T. gondii* and *N. caninum* (Behnke *et al.*, 2015; Lorenzi *et al.*, 2016).

| **Difference** | ***T. gondii*** | ***N. caninum*** |
|---|---|---|
| **Definitive host** | Felids (Cats) | Canine (Dogs) |
| **Intermediate host** | Warm-blooded vertebrate including humans | Bovine, horses |
| **Host range** | Wide | Limited |
| **Transmission strategy** | Mostly horizontally | Mostly vertically |
| **Virulence to hosts** | Strain-specific virulence | Avirulent |
| **Diseases associated with infection in intermediate hosts** | Yes | Yes |
| **Diseases associated with human infection** | Yes | Unlikely |
| **Diseases** | Toxoplasmosis | Neosporosis |

### 1.8.2 Genome structure and genetic differences

Advanced genomic sequencing methodologies have dramatically contributed to our understanding of the differences between *N. caninum* and *T. gondii.* More specifically, the genomic features that underlie biological process, host restrictions and difference in virulence. Recent work has demonstrated that there are significant variations at the genomic and transcriptomic level between *N. caninum* and *T. gondi,* not only between the species but also within their distinct strains. The first whole genome sequence of *T. gondii* was published in 2002 of a type II strain (ME49) with total genomic size 65Mb comprising 14 chromosomes by using Sanger sequencing technology (Tenter, Heckeroth and Weiss, 2000; Dardé, 2004a; Saeij, Boyle and Boothroyd, 2005; Dubey and Jones, 2008; Khan, Dubey, *et al.*, 2011; Mercier and Cesbron-Delauw, 2015).

Subsequently, the genome of *T. gondii* strain ME49 was re-sequenced using NGS to generate a higher quality genome by improving the coverage (Bontell *et al.*, 2009; Khan, Miller, *et al.*, 2011a; Buermans and Den Dunnen, 2014). The first draft of *N. caninum* was published in 2008. The NC-Liverpool strain of *N. caninum* was sequenced to 8 fold depth, assembled into 585 contigs with an N50 of 359 kb totalling 61 Mbp in length. This strain was the first strain isolated in Europe from a 5-week old pup that had symptoms neosporosis disease. The genomes of *N. caninum* and *T. gondii*, are represented by 7266 and 8920 genes respectively distributed in 14 chromosomes. The genes of *N. caninum* strain Liverpool were 7112 protein coding,144 tRNA genes and 10 pseudogenes with the average percentage of GC and AT nucleotides 54.8% and 42.5% respectively. A higher number of genes was identified in *T. gondii* strain ME49 with 8322 protein coding,174 tRNA and 424 rRNA genes. The apicoplast and mitochondrial genomes have also been sequenced in *T. gondii.* The annotations for both species available in public databases (Gajria *et al.*, 2008) Table 1.3 and Table 1.4.

**Table 1.3:** The five references genome strain of *Toxoplasma gondii* and *Neospora caninum.* All the information about the sequences are available in (http://www.toxodb.org).

| Reference strain | Data set source | Coverage |
|---|---|---|
| ***T. gondii* strain *RH* (type I)** | Sanger technology (Wellcome Trust Sanger Institute) (Khan *et al.*, 2006b) | - |
| ***T. gondii* strain *GT1* (type I)** | Sanger and Illumina technology J. Craig Venter Institute (JCVI) and Institute for Genomic Research (TIGR) (Khan *et al.*, 2005) | 67.44 X |
| ***T. gondii* strain *ME49* (type II)** | Sanger technology & Paired – end 454 (Lis Caler; J. Craig Venter Institute (JCVI) (Khan *et al.*, 2005) | 10 X, 26.55 X |
| ***T. gondii* strain *VEG* (type III)** | Sanger technology &Paired-end Illumina sequencing technology J. Craig Venter Institute (JCVI) (Khan *et al.*, 2005) | 77.38 X |
| ***N. caninum* strain Liverpool** | Sanger technology (Wellcome Trust Sanger Institute) (Reid *et al.*, 2012) | 8 X |

**Table 1.4:** The summary of genome features for *T. gondii* ME49 and *N.C*- Liverpool. All data statistics were available in NCBI at http:// https://www.ncbi.nlm.nih.gov/ and from http://toxodb.org/toxo/.

| Genome feature | *T. gondii (ME49)* | *N. caninum (Liverpool)* |
|---|---|---|
| **Estimated Size** | ~65Mb | ~62Mb |
| **Number of chromosomes** | 14 | 14 |
| **Assembly length without sequencing gaps (bp)** | 65,668,120 | 57,524,119 |
| **Number of scaffolds** | 2,277 | 14 |
| **GC content** | 52.28% | 54.8% |
| **Scaffolds N50 (bp)** | 4,973,582 | 5,490,906 |
| **Number of contigs** | 2,508 | 247 |
| **Contig N50** | 1,219,553 | 405,161 |

### 1.8.3 Previous comparative genomic and transcriptome analysis of *T. gondii* and *N. caninum* parasites

Investigations have been published that emphasise how comparative genomic and functional approaches have identified significant changes in evolution and adaptations within Coccidian parasites in general and in the two closely related species *T. gondii* and *N. caninum* in particular (Wasmuth *et al.*, 2009; Behnke *et al.*, 2011; Reid *et al.*, 2012; Adomako-Ankomah *et al.*, 2014). A total of 87,736 protein sequences were identified from 15 apicomplexans revealed a large expansion of specific genes within species that distinguishing this phylum's members. It has been noticed that most of these protein families were associated with the host - parasite interaction and survival strategy (Wasmuth *et al.*, 2009). One of the key features of pathogen genomes are clustering of variable genes, often in telomeric and subtelomeric locations. Cluster of genes share similar sequences and are likely to serve similar functions. Telomeric and sub telomeric positions also facilitate genome rearrangements to create diversity (Barry *et al.*, 2003). However, in some species gene families have expanded in non-telomeric regions. An example is the ApiAP2 gene family, where gene products are associated with virulence and which are randomly distributed across chromosomes (Kissinger and DeBarry, 2011; Reid *et al.*, 2012; DeBarry and Kissinger, 2014).

An example of the expansion of gene families was noticed in *Trypanposoma burcei* for the gene family known as variant surface glycoproteins (VSGs), where only 5% of the VSGs were functional and the rest were categorized as pseudogenes. These densely packed genes have varying clusters which play a key role in manipulation of the host immunity system and regulation of antigenic variation which changes the trypanosome surface (Barry *et al.*, 2005; Berriman *et al.*, 2005; El-Sayed *et al.*, 2005; Jackson *et al.*, 2012; Forrester and Hall, 2014; Mugnier, Stebbins and Papavasiliou, 2016). Further examples of gene family expansion was observed in *Plasmodium falciparum* for the VARs family which consists of around 60 genes per strain (Gardner *et al.*, 2002; Hall and Carlton, 2005).

It has been documented that there was association between VAR genes and regulation of antigenic variation, located towards the telomeric regions. The VAR genes code for proteins are important virulence factors, which mainly involved in host-parasite interaction and contribute to the development of malaria (Wellems, Hayton and Fairhurst, 2009). It is tempting to speculate that some of the divergence between the two closely related organisms may be due to the expansion of these multigene families that are primarily located in sub telomeric regions such as Surface antigen family (SRSs) and those for apical proteins that are released from rhoptry (ROPs), micronemes (MICs) and dense granule (GRAs) organelles, that are all involved in host parasite interactions (Cesbron-Delauw, 1994; Ferguson *et al.*, 1999; Lekutis *et al.*, 2001; Sinai and Joiner, 2001; Jung, Lee and Grigg, 2004; Tonkin *et al.*, 2010; Reid *et al.*, 2012; Wasmuth *et al.*, 2012; Macêdo *et al.*, 2013; Jensen *et al.*, 2015). Reid *et al*., 2012, identified that the genome content of *T. gondii* and *N. caninum* was more than 90% similar. Despite this similarity, these organisms differ dramatically in their repertoires of species - specific genes.

However, a large proportion of species-specific genes were found in *T. gondii* and *N. caninum* totalling 39.1% and other the genes, considered as apicomplexan genes accounted for 28.5%. Based on the transcriptome analysis of the two species, a total of 231 and 113 species-specific genes were identified with no orthologues or paralogues in *T. gondii* and *N. caninum* respectively (Reid *et al.*, 2012). In a further comprehensive study of the evolution of the genomes these two closely related coccidian species, the evidence for novel genes was confirmed by transcriptomic and proteomic data from the tachyzoites. The gene products contributed to disease and the biological mechanism of the host cell invasion. These results have improved the current genome annotations of *T. gondii* strain VEG and *N. caninum* strain Liverpool, due to corrections that have been made in over a third of previously annotated gene models and new annotation of more than half the proteins in untranslated regions (UTRs) that might associate to regulate transcription (Ramaprasad *et al.*, 2015). By comparing data from different next generation sequencing platforms, gene duplications were identified across the recent annotated assemblies of these closely related species (DeBarry and Kissinger, 2014).

Such duplications or deletions are essential sources of genetic diversity in *N. caninum* and *T. gondii* strains and within members of the same clonal lineage (Adomako-Ankomah *et al.*, 2014; Cheng *et al.*, 2015; Lorenzi *et al.*, 2016). A recent comparative genomic analysis of three coccidian species (*T. gondii*, *N. caninum* and *H. hammondi* ) revealed that there was significant enrichment of gene families including surface antigen proteins (SRSs), rhoptry genes (ROPs)micronemes (MICs) and dense granules (GRAs). One gene family has been characterised in *T. gondii* known as the Toxoplasma gondii gene family (TgFAM) which contains five sub families (A-E), which was previously reported as the Toxoplasma-specific family (TSF). Some of those expanded TgFAM genes were clustered in telomeric regions, while other had copy number variations that caused different phenotypic features (Reid *et al.*, 2012; Lorenzi *et al.*, 2016).

### 1.8.3.1 Surface antigens proteins

The surface of protozoan parasites *T. gondii* and *N. caninum* are coated with an array of glycosylphosphatidylinositol (GPI) anchored antigens termed surface antigens (SAGs), members if the SAG1- related sequences superfamily of proteins (SRSs). The SRS family has been implicated in immune modulation, host – parasite interaction, initial attachment of tachyzoites to the host cell surface and regulation of the virulence. Significant expansion of SRS genes in both parasites was due to increased rearrangement and recombination rates, more significantly in *N. caninum* than *T. gondii* (Risco-Castillo *et al.*, 2011; Reid *et al.*, 2012; Wasmuth *et al.*, 2012; Adomako-Ankomah *et al.*, 2014; Cheng *et al.*, 2015; Reid, 2015; Lorenzi *et al.*, 2016; Bezerra *et al.*, 2017). This suggested the important link of expansion of SRS that rapidly evolved, introducing phenotypic differences significant in the pathogenicity of both coccidian species (Risco-Castillo *et al.*, 2011; Wasmuth *et al.*, 2012; Cheng *et al.*, 2015; Bezerra *et al.*, 2017). It has been demonstrated that some SRS proteins are differentially regulated, expressed at different developmental life stages. Stage specific expression of SRSs has been identified in each strains of *T. gondii* and in *N. caninum* respectively.

An example of tachyzoite-specific proteins includes; SRS29B, SRS34A, SRS29C and SRS57 that appear in a strain type dependent manner resulting in different levels of virulence among strains of *T. gondii* (Niehus *et al.*, 2014). Bradyzoite specific antigens include SRS9 and BAG1 that have been linked to altered parasite replication within host cells and persistence of a chronic infection in the host tissue (Jung, Lee and Grigg, 2004; Kim and Boothroyd, 2005; Kim, Karasov and Boothroyd, 2007; Sullivan and Jeffers, 2012). Two more tachyzoite- specific proteins are SAG1 and SRS2 that are involved in adhesion and penetration the host cell (Hemphill and Gottstein, 1996). It was widely believed that the expansion of SRS genes was significantly higher in *T. gondii* strain ME49 than in *N. caninum* strain Liverpool, however Reid and his colleagues (2012) noticed for the first time that *N. caninum* has more SRSs than *T. gondii,* which might support the theory of limited of host range in *N. caninum* during the rapidly growing tachyzoite stage (Reid *et al.*, 2012).

As a component of the parasite surface, a highly polymorphic family of 26 and 38 GPI-anchored proteins were identified in both *T. gondii* and *N. caninum* genomes respectively. This family known as SAG-unrelated surface antigens (SUSA) are found in tandem arrays and clustered in chromosomes VI, IX and XII with a high level of polymorphisms. SUSA1 and SUSA2, two members of this family, were found on the parasite surface and expressed in the bradyzoite stage in *T. gondii* (Pollard *et al.*, 2008; Reid *et al.*, 2012; Reid, 2015). Taken together, the various members of the SUSA family are likely essential for attachment and invasion of host cells, specifically SUSA1 which is among to the bradyzoite - specific antigens indicating the potential role of this antigen in chronic infection. Functional analyses revealed that the SUSA family share similar functions and locations to the SAG family (Pollard *et al.*, 2008).

#### 1.8.3.2 Secreted encoded proteins

Recent comparative analysis reported significant differences in the secreted proteins between the two parasites including ROPs, MICs, GRAs and other gene families (DeBarry and Kissinger, 2014; Lorenzi *et al.*, 2016). There was evidence that the secreted proteins likely play an important function in the coccidian parasites and potentiality responsible for the pathogenesis, host range specificity, adaptations and genetic diversity in the gene families. The host – parasite interactions have been comprehensively investigated in *T. gondii* however limited information is known about *N. caninum* (Wasmuth *et al.*, 2009, 2012; Al-Qassab, Reichel and Ellis, 2010; Reid *et al.*, 2012; Lei *et al.*, 2014; Goodswen *et al.*, 2015; Reid, 2015; Lorenzi *et al.*, 2016). As mention in section 1.6.1, recent genomic analyses of *T. gondii* and *N. caninum* identified divergence of virulence causative genes between the two closely related organisms. There was a significant reduction of virulence in *N. caninum* due to pseudogenisation of the ROP18 gene (Reid *et al.*, 2012; Lei *et al.*, 2014). This gene (TgROP18) has been discovered recently in *T. gondii* and is considered a key virulence effector causing a high virulence by phosphorylation of the host cell and protection against the immune response. Conversely, *N. caninum* was unable to phosphorylated its hosts - immunity-related GTPases, which was confirmed to be key to virulence in *T. gondii* (El Hajj *et al.*, 2007; Niedelman *et al.*, 2012; Reid *et al.*, 2012; Lei *et al.*, 2014; Behnke *et al.*, 2015; Wang *et al.*, 2017).

Remarkably, it was observed there were a divergence in other apical genes including GRA11, GRA12, ROP2A, ROP2B and ROP8, which were missing from the *N. caninum* genome completely (Sinai and Joiner, 2001; Michelin *et al.*, 2009; Reid *et al.*, 2012; Talevich and Kannan, 2013b). Large numbers of genes have an orthologues between the two species, as noticed in *N. caninum* that has clear orthologous of multiple virulence factors including ROP5 and ROP16 in *T. gondii* (Ong, Reese and Boothroyd, 2010; Shwab *et al.*, 2016; Ma *et al.*, 2017a).

### 1.9 Strain variations of *T. gondii*

Several investigations of genotyping markers have been performed to determine the genetic diversity and the population structure of *T. gondii* strains using different approaches to be genotyping the recombinants (three clonal types) as well as atypical or exotic strains that have novel polymorphisms or new alleles. One of the systems used for epidemiological and genetic diversity analysis was the multiloucs enzyme electrophoresis (MLEE) technique. A total of 7 isolates from France fell typed that into three zymodemes (Z1, Z2 and Z3) by using 4 markers (Darde, Bouteille and Pestre-Alexandre, 1988). Later, another study used the same technique confirmed that there was genetic diversity between 35 isolates of *T. gondii* collected from North America and Europe. These were classified into five main zymodemes (Z1-Z5) by using 6 markers (Dardé *et al.*, 1992). Howe and Sibley (1995) reported that there were three main lineages namely types I, II and III from animals and humans. This was determined by using the restriction length polymorphism (RFLP) method to examine a total of 106 strains at 6 loci (Howe and Sibley, 1995a). Microsatellite (MS) genotyping method was used for some reference strains of *T. gondii* through the application of 5 markers that genotyped 43 isolates collected from Africa, North America, South America and Europe in agreement with the Howe and Sibley (1995) findings (Ajzenberg *et al.*, 2004).

The above methods for genotyping different strains of *T. gondii* requires a high number of purified parasites and is highly influenced by contaminations with host DNA. Another genotyping method PCR-restriction fragment length polymorphism (RFLP) followed by multi-locus sequencing typing (MLST), has been used widely with isolates from different geographical regions ( Sibley *et al.*, 2009).

The first study using the PCR-RFLP method was reported by Howe *et al*, (1997) using the SAG2 marker for genotyping a total of 68 isolates from France. Further investigations based on these methods were performed in different countries using this single marker to differentiate the three major groups from each other, as well as identity the presence of atypical and recombinant strains (Sibley and Boothroyd, 1992; Howe and Sibley, 1995; Howe *et al.*, 1997; Lehmann *et al.*, 2000; Fuentes *et al.*, 2001; Su *et al.*, 2004; de Melo Ferreira *et al.*, 2006; Su, Zhang and Dubey, 2006). However, PCR-RFLP was not capable of discriminating the three major groups (I, II and III) and the recombinant and exotic (atypical) strains using a single marker. An improvement came from application of multi-locus sequencing typing (MLST) to get accurate detection of the true level of diversity by identifying single nucleotide of polymorphism (SNPs) between the strains of *T. gondii* especially, from new strains (Sibley *et al.*, 2009).

Studies have increasingly used the MLST method due to the high resolution to detect mutations in the coding regions of *T. gondii* strains (Lehmann *et al.*, 2000; M. E. Grigg *et al.*, 2001; Michael E. Grigg *et al.*, 2001; Frazão-Teixeira *et al.*, 2011; Khan, Dubey, *et al.*, 2011; Khan, Miller, *et al.*, 2011b). New advances is whole genome sequencing and phylogenetic analyses revealed a much more complex population structure with highly diverged sequences which included new alleles. These new groups of strains were classified as atypical (exotic) strains and are largely distributed in South America (Lehmann *et al.*, 2004; Sibley *et al.*, 2009; Su *et al.*, 2012). Analyses of the major lineages in animals circulating in North America and Europe showed strains mainly belonged to type II (Boothroyd and Grigg, 2002; Ajzenberg *et al.*, 2004). Further study has shown that most cases in AIDS patients and congenital infectious in North America and Europe also belonged to type II. In France, more than 90% of isolates were classified as type II from both domestic animals and humans (Tenter, Heckeroth and Weiss, 2000).

The other two types (I and III) were also found in European countries. Type III was the predominant type most frequently encountered in Southern Europe (Howe and Sibley, 1995b; Daniel K. Howe *et al.*, 1997; M.-L. Dardé, 2004a). It was obvious from the above genotyping methods that utilising limited number of loci gave insufficient resolution for discovering the divergence between different linages and within strains

in addition to the high cost of the genotyping methods (Depristo *et al.*, 2011; Yu and Sun, 2013). In order to understand the population structure of *T. gondii* at greater depth sequencing is required to demonstrate polymorphisms in the DNA sequences of the different strains of *T. gondii* across the genomes (Ellegren and Galtier, 2016). The power of this technology is from identification of a much larger number of variations between strains compared to the reference genomes of the *T. gondii* parasite.

More importantly, it can give information on pathogenic isolates that have highly polymorphic genes across the whole genome and transcriptome provides the basis for epidemiological investigations (Bontell *et al.*, 2009; Wang *et al.*, 2012; Buermans and Den Dunnen, 2014; Khan, Shaik, *et al.*, 2014; Behnke *et al.*, 2015; Cheng *et al.*, 2015; Lorenzi *et al.*, 2016). Recently, by using NGS, which is now considered the gold standard method through generating large amount of data from RNA-Seq and DNA-Seq of different isolates, hotspots mutation have been detected that confer the phenotypes and yielded a wide range of information related to evolution, transmission, genetic diversity and virulence (Sidik, Huet and Lourido, 2018).

However, one of the main challenges of this approach is sequencing errors from different sources; mapping, repetitive sequences and duplications, depending on the platform used (Rizzo and Buck, 2012; Chen *et al.*, 2013; Treangen and Salzberg, 2013; Buermans and Den Dunnen, 2014; Ari and Arikan, 2016; Abnizova, Boekhorst and Orlov, 2017). It has been found that there are major strain haplotypes, which cluster different genotypes into major groups to define the population structure of *T. gondii* (Khan *et al.*, 2007; Khan *et al.*, 2011; Su *et al.*, 2012; Lorenzi *et al.*, 2016). A total of 46 strains collected from Europe, North America and South America grouped into 11 distinct haplogroups based on intron sequences of from different loci 1,2 and 3 haplogroups mainly corresponded to the three clonal lineages I, II and III that were located in North America and Europe. However the remaining groups predominated in South America and haplogroup 6 was widespread thus reflecting geographical separation between strains of *T. gondii* (Khan *et al.*, 2007).

Another comprehensive study of the population structure of *T. gondii* revealed that a total of 138 unique genotypes grouped into 15 haplogroups and clustered into six clades: Haplogroup1 (HG1), 6 and 14 formed clade A; HG 4 and 8 defined clade B; HG3 formed clade C; HG 2 and12 formed clade D; HG 9 formed clade E and HG 5, 10 and 15 formed clade F based on three types of markers; microsatellite, RFLP and intron sequences analyses, which confirmed the previous findings of Khan *et al*., (2007). This included demonetisation South America strains showed a high level of divergence (Minot *et al.*, 2012; Su *et al.*, 2012).

A recent population multi-isolate genetic study of *T. gondii* using 62 isolates collected from different geographical regions used genome-wide SNP analysis revealed that there was a large number of conserved haploblocks that grouped into clade-specific clusters contained groups of gene families including SRS, MIC, GRA, ROP and TgFAM proteins that significantly influenced the population structure of *T. gondii.* Furthermore, tandem amplification and diversification of the proteins clustered in gene families is the primary characterisation that distinguishes the different genomes of these biologically diverse isolates that influence host-parasite interaction (Lorenzi *et al.*, 2016). Furthermore, Comparison of population structure analyses of isolates from animals showed there were widespread clonal strains in Europe, North America and Asia, but there was a high degree of diversity was seen in some regions, namely Central America, South America and Africa (Khan, Ajzenberg, *et al.*, 2014; Sharif *et al.*, 2017).

A recent genetic diversity study (Sharif *et al.*, 2017) of a collection of *T. gondii* isolates to assess and compare the different types in ruminants found that the most prevalent strains were type II (81.4%). Additional atypical or exotic genotypes were identified totalling 82 out of 215 (38.13%) associated with geographical distribution of the host. There were atypical strains including Chinese 1, types Br (I, II, III and IV), and type 12 identified. Further genetic investigations have concluded that a total of 1475 isolates could be closed into 189 genotypes from sheep, goats, camels and cattle. Types1, 2 and 3 were distributed globally, but there were new genotypes totalling 646 representing 156 genotypes in Central and South America. All the genotyping diversity results of different samples from varied geographical regions confirmed there

was a high genetic diversity in South and Central America (Shwab *et al.*, 2014). A possible explanation for this expansion in South and Central America is human mass migrations including agricultural activities, human exchange and trading goods.

Another possible explanation for the genetic diversity is the potential long-distance migration of animal species such as cats dogs, rats, mice and other animals or by accidental transport of infected animals (Sibley *et al.*, 2009; Khan, Ajzenberg, *et al.*, 2014; Shwab *et al.*, 2014). It is worth noting that there was an association between the infection with the number of hosts across several populations and how they adapt to the different environments. More specifically, warmer climates might allow a higher number of the agent's oocysts of infection to survive for a long time and a higher prevalence of infection was classically noticed for warm countries. As discussed earlier in this chapter (Section 1.4), one mark of the success of *T. gondii* is the highly resistant oocyst form that is transmitted through water contaminated with cats faeces (Frenkel, Dubey and Miller, 1970; Dubey, 1998).

## 1.10 Strains variations of *N. caninum*

Previously, genetic heterogeneity and biological diversity of *N. caninum* strains have been analysed to investigate the intra- species diversity from different hosts and from worldwide origins. These finding from different molecular methods including PCR-based techniques, 18S-like ribosomal DNA (small subunit-rDNA), Single nucleotide polymorphism (SNPs), Microsatellites, Internal transcribe4d spacer sequences (ITS1) and Rapid amplification of polymorphic DNA (RRAPD-PCR) have been broadly applied to investigate the significant biological and genetic discrepancy between the distinct strains and determine the relatedness between the geographical distribution and the population structure of *N. caninum* isolates (Schock *et al.*, 2001; Al-Qassab *et al.*, 2009; Al-Qassab, Reichel and Ellis, 2010; Regidor-Cerrillo *et al.*, 2013). Understanding what functional factors dominate in structuring the genetic diversity and the patterns of polymorphisms across the genomes were explained the influence of those forces on the adaptive evolution, conservation within genomes, host preference, virulence and growth rate (Schock *et al.*, 2001; Regidor-Cerrillo *et al.*, 2006; Al-Bajalan *et al.*, 2017). A total of 108 clinical samples that were collected from bovine and ovine from four different origins countries, Scotland, Germany, Spain and Argentina. The result support a close relationship between Spanish and Argentinean populations and the Sottish and German population. Notably, there were some factors other than geographical separation that influence the level of the genetic diversity between the populations (Regidor-Cerrillo *et al.*, 2013).

Several factors that might have influence on the virulence includes the parasite life stage used to initiate infection, route of inoculation, inoculum dose, growth rate and other biological characteristics in cell culture derived parasites. In addition, the association between the genetic diversity of *N. caninum* isolates and their pathogenicity has not been conclusively demonstrated and it stills unclear if the genetic structure of the parasite is a determinant of host clinical manifestations (Dubey, Schares and Ortega-Mora, 2007b; Al-Qassab, Reichel and Ellis, 2010). As to date there have been several investigations on the differences in virulence among *N. caninum* isolates that collected from different hosts (Lindsay and Dubey, 1989; Anderson *et al.*, 1992; Holmdahl *et al.*, 1995; Barber and Trees, 1998; Atkinson *et al.*, 1999).

The experimental infection of different hosts was confirmed that *NC-1* and *NC-Liverpool* strains appeared that were highly pathogenic than other *N. caninum* strains (Collantes-Fernández *et al.*, 2006). With the exception of *NC-Bahia* from Brazil, this isolate was less pathogenic than other isolates and supported the hypothesis of differences in the pathogenies rate was related to the replication speed of the tachyzoites among strains in tissue under identical tissue culture conditions (Luis F P Gondim *et al.*, 2004). *NC-Bahia* isolates take the half of time to destroy the Vero cell monolayer than *NC1* strain. Recently, no clustering was found among strains that related to the host or the geographical origins from six *N. caninum* strains that isolated from canine and bovine showed complete sequence conservation (Schock *et al.*, 2001). Further evidence of no geographical segregation between N. caninum isolates was reported by Calarco et al., (2018), which confirmed the hypothesis of the genetic diversity among isolates of *N. caninum* that likely attributed to the differences in the sub set of genes that involved in virulence between the strains that grouped them into two main clades (Calarco, Barratt and Ellis, 2018).

## 1.11 Genome sequencing

Sanger sequencing is the first-generation DNA sequencing technique. It was developed by Frederick Sanger and colleagues in 1977 and is still a gold standard of sequencing technology as served in the 1000 Human Genomes Project (Singh, 2017). This technique depends on chain termination. Radioactively or fluorescently labelled single nucleotides are added by DNA polymerase. Although the radioactive labelling and autoradiography were initially used for visualising DNA sequence, the fluorescent dye- terminator sequencing is now the mainstay in automated sequencing owing to its greater expediency and speed. In this modified method, the emitted lights are excited by laser in the DNA sequencer while passing through the detection region. The DNA sequence is revealed from the order of the fluorescent fragments (Schuster, 2008; Estrada-Rivadeneyra, 2017). Sanger sequencing bears technical limitation for sequencing large number of targets as it is not time or cost efficient. It requires PCR amplification for DNA template (library) preparation which usually involves either DNA cloning or gel purification. The estimated cost for an 800 bp sequencing reaction is £3. Another big challenge for this method is that it can directly sequence only relatively short (300 - 1000 nucleotides) DNA fragments in a single reaction. All of the above limitations hamper its further application and justify development of new techniques with a higher capacity and efficiency to sequence a big size of DNA region in a large number of samples concurrently (Forrester and Hall, 2014).

Next generation sequencing (NGS) is a new technique developed only a decade ago (Ari and Arikan, 2016; Abnizova, Boekhorst and Orlov, 2017; Singh, 2017). It is characterised by massively parallel sequencing in which up to millions of DNA fragments from a single sample or multiple samples are sequenced simultaneously. Unlike Sanger sequencing, this technology has witnessed improvement in mutation detection at higher sensitivity based on the large order of sequencing magnitudes each target nucleic acid can achieve through deep sequencing approaches (Nowrousian, 2010; Depristo *et al.*, 2011). Different NGS platforms are commercially available and newly emerging platforms are continuing to be developed. These platforms have different throughput capacity. This has made sequencing facilities accessible to more labs and highlighted an interest for their use in a clinical setting which require cheaper, faster, and easier-to-use sequencer (Ajay *et al.*, 2011).

The use of next generation sequencing method (NGS) has expanded the knowledge of the genomic research in infection diseases such as Toxoplasmosis and Neosporosis as it has led to discovery of new recurrently mutated genes which were previously unknown in both species (Reid *et al.*, 2012; Adomako-Ankomah *et al.*, 2014; English, Adomako-Ankomah and Boyle, 2015). Functional and clinical studies of these novel gene mutations have discovered new mechanisms implicated in the pathogenesis of the disease, revealed new insights into parasites molecular evolution that could ultimately translate into improvements in the clinical management of patients.

## 1.12 Genome Assembly

With the recent advanced next-generation sequencing (NGS) methodologies it has become much more affordable to assemble the high-throughput RNA and DNA-sequencing data of most apicomplexan parasites including *T. gondii* and *N. caninum* genomes (Church *et al.*, 2015; Chen *et al.*, 2018). Most recently, the number of available assembly algorithm tools was dramatically increased. There are many consideration should be taken when using suitable tool including time, computational resources, accuracy and memory (Phillippy, 2017). There are two different types of sequence assembly including mapping assembly and *De novo* assembly. The mapping assembly refers to sequencing the reads to the known reference genome however the *De novo* assembly refers to use the sequencing reads of the genomes or transcriptomes when there is no known reference genome to mapped the reads that assembled them into contigs (Kitts, 2003; Hassan *et al.*, 2012).

## 1.13 Single nucleotides polymorphisms detection and analysis

Recently, re-sequence different genomes at high coverage is an important factor in Single nucleotides polymorphisms (SNPs) detection from several next-generation sequencing platforms. As we mention before in section 1.9, the power of the advanced sequencing approaches is to generate millions of reads to identify a wide range of genetic variants by mapping the sequence reads to the known reference genomes (Nowrousian, 2010; Bauer, 2011; Depristo *et al.*, 2011). For the SNP detection from the two organisms, there are many variants callers used to determine the potential hotspots that that increase susceptibility to Toxoplasmosis and Neosporosis diseases (Lorenzi *et al.*, 2016; Calarco, Barratt and Ellis, 2018). By expanding the bioinformatic sources and tools there are many software used to study the SNPs positions, alleles frequencies, structural and functional impacts, the link between the SNPs with the complex disease and more importantly the genes that contribute to host - parasite interaction and virulence. SNPs from different chromosomal regions (Dubremetz and Lebrun, 2012).

According to The Red Queen hypothesis (Paterson, Vogwill, Buckling, Benmayor, Andrew J. Spiers, *et al.*, 2010), there is a dramatic evolution in the genes that are implicated in the host – parasites interaction such as those seen between many apicomplexan species and their hosts. Interestingly, antagonistic coevolution can drive a significant increase in the rate of molecular evolution between species and generate a high degree of genetic divergence among genomic and phenotypic changes, virulence, adaptation and host defence mechanisms (Ellegren and Galtier, 2016). A newer computer programs and caller algorithms are increased to solve some technical problems such as increased number of false-positives, false-negative variant calls, repetitive reads, sequencing errors to detect accurate and confidence SNPs from NGS data (Yu and Sun, 2013). Many applications have been designed to identify the SNPs including GATK, SAMtools, Bcftools,Varscan, Platypus, FreeBayes and VarDict, LoFreq (Sandmann *et al.*, 2017).

## 1.14 Copy number of variation detection and analysis

Another type of genetic variations is the copy number of variations (CNVs) that distributed across the genome. The importance of CNVs is associated with particular disease and host-parasite interaction (Barry *et al.*, 2003; Xiao *et al.*, 2019). The advantages of NGS technologies are to provide the high coverage to detect accurate CNVs more specifically the novel CNVs that involved in different phenotypes such as virulence (Behnke *et al.*, 2011). The gain and loss in the sequences of different genomes are duplication and deletion that are directly influence gene dosage(Zhou *et al.*, 2011). Recently, the duplication is connected to gene families which includes virulence genes and having CNV in different copies in many apicomplexan parasites such as *T. gondii* and *N. caninum* (Reid, 2015).

There are many strategies used for identifying CNV including read depth (RD), split read (SR), *do novo* assembly of genome (AS), paired end mapping (PEM) and the combination of different approaches (CB) each approach has advantages and limitations depending on the NGS data, input data, sensitivity and specificity of the tool used. One of the main challenging to find accurate CNVs is the low coverage of the reads generated that positively correlated with the low number of CNVs (Cingolani *et al.*, 2012; Xiao *et al.*, 2019).

## 1.15 Aims of the thesis

Recent comparative genomic analyses of *T. gondii* and the closely related *N. caninum* parasites have identified a set of species - specific genes in *T. gondii* and *N. caninum* (Reid *et al.*, 2012; Lorenzi *et al.*, 2016). These species-specific genes are largely hypothetical proteins of unknown function. However, several investigations of other parasites with species-specific genes have shown they often associated with host-pathogen interaction. We hypothesize that these genes play important roles in mechanisms of infections for a wide range of vertebrate hosts. There are still many unanswered questions about the host range, host parasite interface, the effective transmission mode and the evolutionary mechanisms. This includes questions such as why *T. gondii* infects humans, but *N. caninum* does not and why *N. caninum* has a more limited host range than *T. gondii*. Selecting a wide range of isolates in both species will help to discover if the variations in their genomes may contributing to phenotypic changes among strains that underlie their divergent species- specific genes affecting their host-parasite mechanisms and pathogenic differences. Multiple comparison of six strains of *T. gondii* (*T. GT1, T. MAS, T. P89, T. COUG, T. VEG* and *T. CAST*) and three *N. caninum* strains (*NC-Bahia, NC-1* and *NC-Liverpool*) isolated from different hosts and geographical locations were used to look at divergence between strains. No data on inter-species comparative genomic and genetic diversity have been reported previously for these species associated with mutation rates and phylogenetic relatedness.

In order to explore these biological characteristics of important veterinary and medical pathogens was used to study the genetic diversity and the divergence rate between the two organisms and their different strains.

In Chapter one, we covered major aspects of *T. gondii* and *N. caninum* (life cycle, life stages, mechanism of infection, differences and similarities, comparative genomic investigations).

In Chapter two, we presented all the methods and materials that were used including parasite isolation, host and parasite culture and bioinformatics tools used to generate analyses of the polymorphism rates after comparing them to the references genomes to find out the genetic diversity background between the strains and then between the two organisms.

In Chapter three we performed a comparative genomic analysis of *T. gondii* and *N. caninum* to identify species-specific genes unique to each species which might play a direct role in divergence between the two species and associate with host - parasite interaction.

In Chapter four, we performed DNA sequencing in different strains of *N. caninum* and highlighted multiple genomic comparison to detect the genetic diversity among them.

In Chapter five we had the same methods and analyses but on the six strains of *T. gondii*.

In Chapter six offers the general discussion and parasites for future work between the two organisms to understand biological and genomic differences.

# Chapter 2: Materials and Methods

## 2.1 Comparative genomic analysis of *T. gondii* and *N. caninum* parasites

### 2.1.1 Analysis of orthologues genes of *T. gondii* and *N. caninum* using OrthoMCL

Protein sequences were downloaded in FASTA format files from the *Toxoplasma* genomic resource ToxoDBv29 database (http://toxodb.org/toxo/) for *T. gondii* strain ME49 and *N. caninum* strain Liverpool. OrthoMCL tool v2.0.3 (http://orthomcl.org/orthomcl/), which relies on Markov Clustering method (MCL) was used to generate clusters of orthologous proteins between species and within species. Default parameters were used as following; E-value threshold of 1e-5 was applied to all - vs - all BLASTP, -a 30 -m 8, -v 1000 and -b 1000. MCL was run using a clustering granularity value of 1.5. The resulting output was assigned into 'orthologous' or 'paralogous' categories according to the following criteria; a cluster was considered 'orthologous' if they contained at least one sequence from *T. gondii* and one sequence in *N. caninum* (two taxa). If the sequence from one taxon (single specie with more than one copy) this, we considered them as paralogous that likely as a result of duplication events. All the remaining sequences that were not belonging to the two previous categories which containing no orthologous or paralogous sequences in any of the two genomes were considered 'unique' with singleton genes, having one copy species-specific sequences. We downloaded the list of species-specific genes for both species from Reid findings then we performed the comparison of the putative species specific genes for *T. gondii* and *N. caninum* from our analysis with those obtained from data published in Reid analysis (Reid *et al.*, 2012).

Next, all the genes found to be unique to *T. gondii* and *N. caninum* based on the OrthoMCL clustering as defined above were grouped for further analysis to identify the biologically important families of proteins; SAG1- related sequences (SRS) genes, Rhoptry (ROPs) genes and dense granules (GRA) genes, Micronemes (MICs) genes, Toxoplasma gondii family proteins (TgFAMS) and other gene families were identified within the annotated genome file of TgME49 and NCLIV from the ToxoDBv29 database (Figure 2.1).

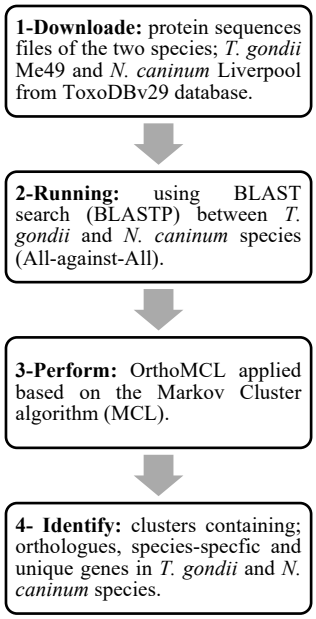



**Figure 2.1**: Schematic representation of the four search strategies used to find orthologues, species – specific and unique genes in *T. gondii* (ME49) and *N. caninum* (Liverpool) with the bioinformatics tools used.

## 2.2 Strains selection

The nine strains of *N. caninum* (three strains only based on the availability from the sources) and six strains of *T. gondii* parasites listed in Table 2.1 analysed in this study were obtained from laboratory-maintained cultures which had been determined to be free from any sources of contamination by checking the viability, cell morphology and purity based on the microbiology workflow and the information in each certificate of analysis per strain provided from American Type Culture Collection (ATCC) https://www.lgcstandards-atcc.org/?geo_country=gb.

These isolates were obtained from experimentally infected hosts provided from the product sheet per strain including animals and humans with congenital toxoplasmosis and AIDS patients. All the isolates were maintained in tissue culture conditions following American Tissue Culture Collection (ATCC) protocols. These strains were selected due to phenotypic variation previously reported in several genetic diversity studies (Lorenzi *et al.*, 2016; Calarco, Barratt and Ellis, 2018) as discussed in depth in Chapter 1 (Section 1.8). The details of isolates and geographic distributions are shown in Table 2.1 and Figure 2.2.

**Table 2.1**: The representative isolates of *T. gondii* and *N. caninum* used in this study.

| Organism | ATCC number | Isolates | Host of origin | Depositor name |
|---|---|---|---|---|
| ***T. gondii*** | ATCC-50853 | *T. GT1* | Skeletal muscle of goat | D Sibley |
| | ATCC-50861 | *T. VEG* | Human with AIDS | LD Sibley |
| | ATCC-50870 | *T. MAS* | Human with congenital toxoplasmosis | LD Sibley |
| | ATCC-50868 | *T. CAST* | Human with AIDS | LD Sibley |
| | ATCC-50879 | *T. P89* | Pig | LD Sibley |
| | ATCC-PRA-356 | *T. COUG* | Cougar | LD Sibley |
| ***N. caninum*** | ATCC-PRA-138 | *NC-Bahia* | Dog | LF Gondim, MM McAllister |
| | ATCC-50845 | *NC-Liverpool* | Five-week-old Boxer puppy, Liverpool, England | LD Sibley |
| | ATCC-50843 | *NC-1* | Dog | LD Sibley |

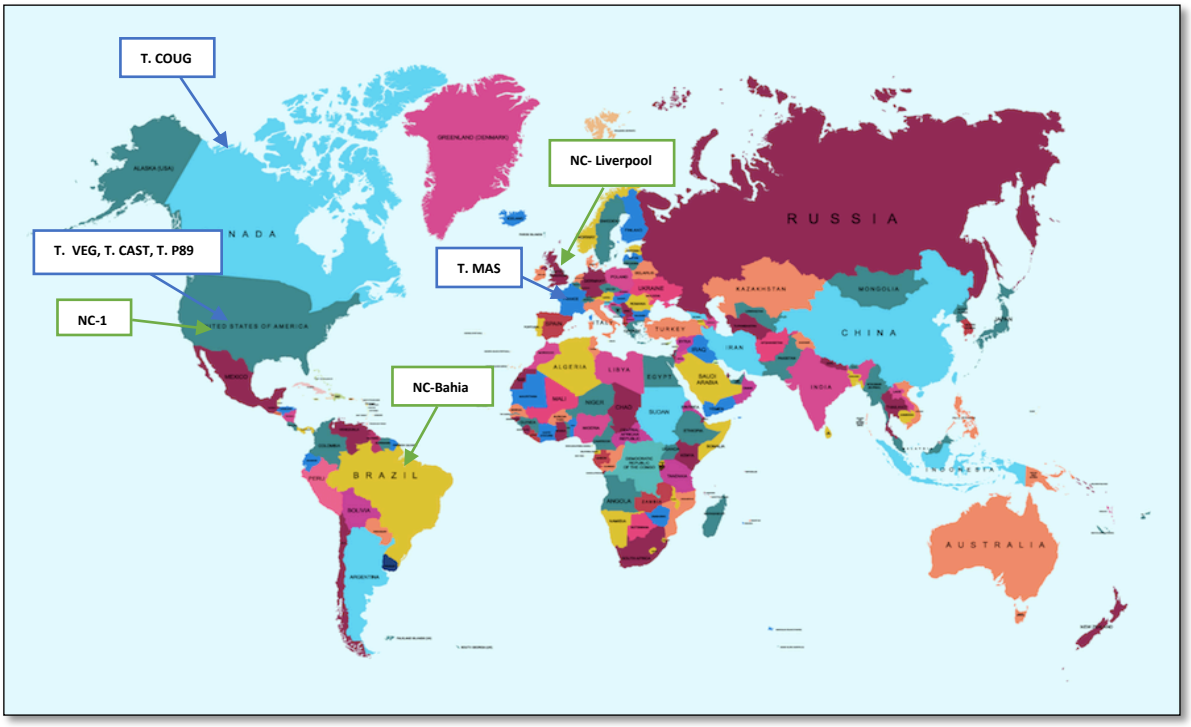



**Figure 2.2:** The geographical origin of the isolates. The blue boxes indicate the *T. gondii* strain locations, and the green boxes indicate the *N. caninum* strain locations.

## 2.3 Cell culture of the hosts and parasites

### 2.3.1 Vero cell passage

The Vero cells line used in this study was isolated from African green monkey kidney cells and were kindly provided from the Institute of Infection and Global Health at University of Liverpool. These acted as host cells for the infective stage tachyzoites in both *N. caninum* and *T. gondii*. Uninfected Vero cells (host cells) were grown in the lab at 37°C in a 5% $CO_2$ humidified incubator in T25s vented cell culture flasks (BD FalconTM) with 5 ml culture media, using filter sterilised RPMI 1640 medium supplemented with 10% FCS and 1% penicillin (sigma-aldrich, 10,000 U/ml). After rinsing with 5 ml PBS buffer solution twice (sigma-aldrich), the cells were then incubated in 1ml trypsin (0.05%) for approximately 5 min and not more than 10 min at 37°C. The cells were flushed down and then re-suspended in 5 ml fresh medium. A haemocytometer was used to determine the number of Vero cells that had been obtained by performing a cell count: X number of cells in 25 squares = $X \times 10^4$ cells $ml^{-1}$, which were then seeded at a ratio of $1 \times 10^5$ per 5 ml RPMI/25 $cm^2$ flask. The cells were ready for passaging after seven days when the parasites starting to egress from the host cell (Vero cell). The host cells were routinely sub cultured every 3 -5 days until the monolayers reached the desired confluence.

### 2.3.2 Parasites passage

To harvest good quality parasites were incubated approximately 24 hours post seeding and after obtaining the desired number of Vero cells, which had reached 10 - 20 % confluence (monolayer of host cell). Different seeds of parasites were counted as described above, inoculated into flasks containing host cells. The Vero cells were infected with $4 \times 10^5$ tachyzoites. *T. gondii* and *N. caninum* tachyzoites were scraped from the cell culture flask using a sterile cell scraper (BD FalconTM) and were separated from the cells by syringing with a small-bore needle (0.7mm) (BD FalconTM). Parasites were then either passage into new cell lines (fresh Vero cells) or harvested for extraction of parasite genomic DNA.

To determine the required number of tachyzoites per millilitre, a Neubauer counting chamber (haemocytometer) was used according to the manufacturer's specifications. Afterwards the number of tachyzoites were calculated and isolated for use in the next round of passaging before centrifugation in 15 ml centrifuge tube at 1500 x g for 10 min. In Figure 2.3 A & B, the host cells are shown before and after infection with the parasites in tissue culture conditions.

## 2.4 Parasite purification

After harvesting the parasites, the tachyzoites were filtered using disposable PD -10 desalting columns with volume 8.3mL (GE Healthcare,17-0851-01). By using the gravity protocol, the tachyzoites were washed twice with 5 ml PBS (pH 7.4) then centrifuged at 1500 x g for 10 min at 4°C. The pellet was re-suspended into 5 ml PBS, the supernatant was discarded, and the final pellet contained the purified tachyzoites. Each pellet contained approximately $1 \times 10^8$ tachyzoites. The parasites were filtered by adding 25 ml PBS to the column. The number of parasites were determined using haemocytometer and the parasites collected in 15 ml centrifuge tube from the column then prepared as starting material for the next step.

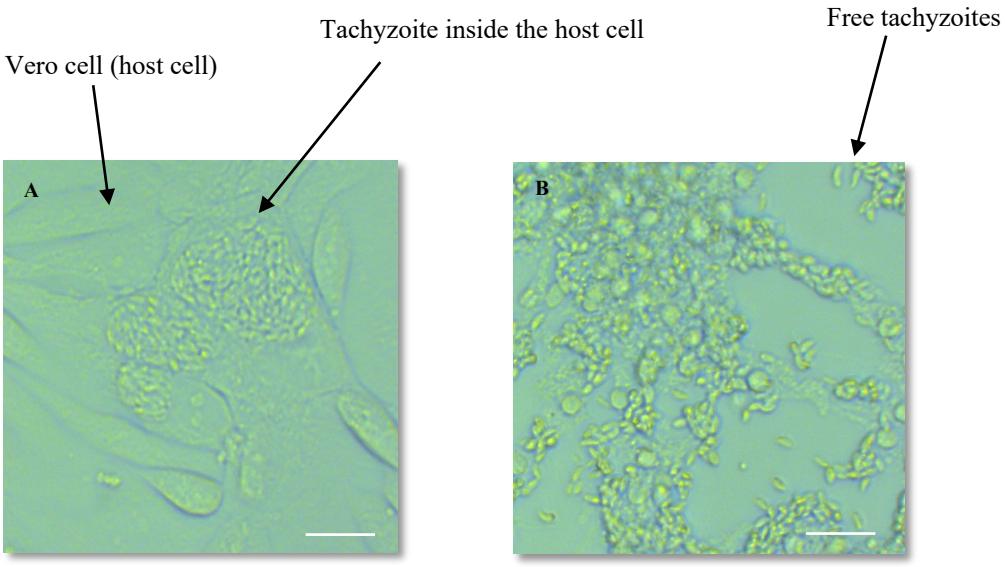



**Figure 2.3: A)** The tachyzoites (infective stage) before passaging inside the Vero cells (host cell); **B)** free after infection ready to passage or harvest. Images were taken from our experiments. Scale bar =10 µm.

## 2.5 Purification of total genomic DNA from tachyzoites

The total genomic DNA was extracted from tachyzoite stage parasites using DNeasy Blood and Tissue Kit (QIAGEN, cat no./ID 69504) https://www.qiagen.com/us/shop/sample-technologies/dna/genomic-dna/dneasy-blood-and-tissue-kit/#orderinginformation. Purification of total DNA was carried out as described in the manufacturer's protocol (spin- column protocol). Firstly, freshly egressed tachyzoites from Vero cells were separated from host cell debris in 15 ml centrifuge tubs with required number of cells depending on the number of parasites collected for 5 min at 3000 *x g*. The pellet resuspends in 200 µl PBS solution. After that, 20 µl proteinase K 40 mAU/mg protein was added into 1.5 ml microcentrifuge tube then added 200 µl Buffer AL (QIAGEN Buffer), without added ethanol, were added and mixed thoroughly by vortexing to yield a homogeneous solution. This was incubated at 56°C for 10 min. Then, 200 µl ethanol (96-100%) was added to the sample and mixed by vortexing. The mixture was then transferred to a 2 ml DNeasy Mini spin column provided in the kit and centrifuged at 6000 g for 1 min.

The flow-through and collection tube were discarded. The DNeasy Mini spin column was placed into a new 2 ml collection tube, 500 µl Buffer AW1 was added, and this was centrifuged (6000 g; 1 min) followed by discarding the flow-through and collection tube. The DNAeasy Mini spin column was transferred to a new 2 ml collection tube, 500 µl Buffer AW2 was added to the sample, and this was centrifuged at 20,000 *x g* for 3 min to dry the membrane of the DNeasy Mini spin column. The flow-through and collection tube were discarded, and the spin column was placed in a clean 1.5 ml or 2 ml microcentrifuge tube. The genomic DNA was then eluted by adding 200 µl Buffer AE directly on the DNeasy membrane, incubating at room temperature for 1 min and centrifuging for 1 min at 6000 g. The quality was checked using A260/280 and A260/230 values from the NanoDrop$^{TM}$ spectrophotmeter and the quantity was checked using the Qubit ds DNA BR Assay Kit (No: Q32850) (Invitrogen). The isolated genomic DNA samples were immediately stored at - 20°C.

## 2.6 Quality control assessment (QC)

All the nine genomic DNA samples were submitted and prepared to meet sequencing requirement by centre for genomic research (CGR). The genomic DNA was suspended in Tris-EDTA buffer (TE) for Illumina DNA fragment libraries. Firstly, QC was determined to ensure all samples had removed contaminants such as RNA, proteins or chemicals that can influence the accuracy of library preparation and the sequencing process. The samples purity was determined using the Bioanalyzer chip (Agilent) to determine RNA integrity number (RIN) values and for overall quality by measuring absorbance and integrity, 260/280 and 260/230 as well as the determining concentrations and volumes were sufficient per sample as starting material to be ready for next stage.

## 2.7 Library Preparation

Libraries were made from samples using two methods by CGR. For all the samples except *NC- Bahia*, the Illumina® TruSeq® Nano DNA Library Prep kit was performed. In brief, the samples were sheared to an average of 350 bp by using a Bioruptor® Pico (diagenode) using predetermined settings. The sheared samples were checked to see the extent of shearing. Then, samples were cleaned with using AMPure XP beads (Agencourt, Protocol 000387v001) to remove any contamination such as residual enzymes, excess primers and salts and can be removed using a simple washing protocol. The resulting purified is essentially free of contaminants to improve the quality of the DNA samples before the next library preparation stage. The end repaired for 30 minutes at 30 °C. Following end repair, a bead-based size selection was performed enrich for 350bp fragments by using Illumina® TruSeq® Nano DNA Library Prep kits. The products were A-tailed with A single 'A' nucleotide by incubating the products at 37°C for 30 minutes to the 3' ends of the blunt fragments to prevent them from ligating to each other during the adapter ligation reaction and ligated to dual ended adapters at 30°C for 10 minutes to ligates multiple indexing adapters to the ends of the DNA fragments, preparing them for hybridization onto a flow cell. The samples were cleaned twice with an equal volume of AMPure beads and amplified for 8 cycles. The products were cleaned and checked for both quantity

with Qubit™ dsDNA HS (Higher sensitivity) Assay Kit (Invitrogen™, Cat. number: Q32854) a DNA HS DNA qubit kit and quality on an Agilent DNA HS bio analyser chip. The sample *NC- Bahia* was converted to a Nextera XT library (Illumina) due to the limited amount of DNA in the sample. Prior to library preparation, the sample was purified with 1.8x volume of Ampure XP beads and the eluted sample was measured by Qubit assay. Tagmentation of input DNA was used and purified following the manufacture's protocol. This was then amplified with 12 cycles of PCR including indexing at this point. The library was purified with Ampure XP and assessed using Qubit assay and Agilent Bioanalyser. All libraries were pooled in equimolar amounts.

Subsequently, a quantitative real-time PCR (qPCR) assay, designed to specifically detect adapter sequences flanking the Illumina libraries was performed using an Illumina® KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, USA). This assay was used to specifically quantify the number of cDNA templates that had both adaptor sequences on either end and therefore those that would successfully form clusters on a flow cell for sequencing. Briefly, a 20 µl PCR reaction (performed in triplicate for each pooled library) was prepared on ice with 12 µl SYBR Green I Master Mix act as dsDNA-binding dye and 4 µl diluted pooled DNA (1:1000 to 1:100,000) depending on the initial concentration determined by the Qubit® dsDNA HS Assay Kit. PCR thermal cycling conditions consisted of initial denaturation at 95°C for 5 minutes, 35 cycles of 95°C for 30 seconds (denaturation) and 60°C for 45 seconds (annealing and extension), melt curve analysis to 95°C (continuous) and cooling at 37°C (LightCycler® LC48011, Roche Diagnostics Ltd, Burgess Hill, UK). The template DNA was denatured for 8 minutes at room temperature using freshly diluted 0.1 M sodium hydroxide (NaOH) and the reaction was subsequently terminated by the addition of 5 µl 0.5M TrisCl PH=8. Following calculation of the molarity using qPCR data, template DNA was diluted to a loading concentration of 300 pM.

## 2.8 Whole genome sequencing (WGS)

The pooled libraries were sequenced using two lanes of the Illumina HiSeq 4000 platform with version 4 chemistry using sequencing by synthesis (SBS) technology to generate paired-end sequencing 2 x 125 bp of 9 indexed libraries. The samples were submitted to the Centre of Genomic Research in University of Liverpool (CGR) for sequencing according to the optimised DNA prep protocol used by the CGR. Figure 2.4 shows the workflow of sequencing steps.

### 2.8.1 Read Processing and quality assessment of the raw sequence data

Raw sequenced data from whole genomic libraries of samples of *N. caninum* and *T. gondii* isolates were obtained from the Center Genomic Research post sequencing from Illumina HiSeq 4000 platform in Fastq.gz formatted files ready for the next downstream analysis. Prior to mapping, the total number of reads were trimmed of matches with adaptor sequences for 3 bp or more by using Cutadapt version 1.2.1 with option -O 3. The presence of the quality score helps trimming or filtering of poor-quality reads. The reads were further trimmed by using Sickle version 1.200 with minimum window quality scores of 20. After quality trimming, reads shorter than 10 bp were removed. The total number of raw and trimmed reads in millions retrieved from all the libraries are illustrated in Figure 2.5. All the sections from 2.6 to 2.8.1 was done by CGR. However, the next flowing workflow steps were done by us in this project.

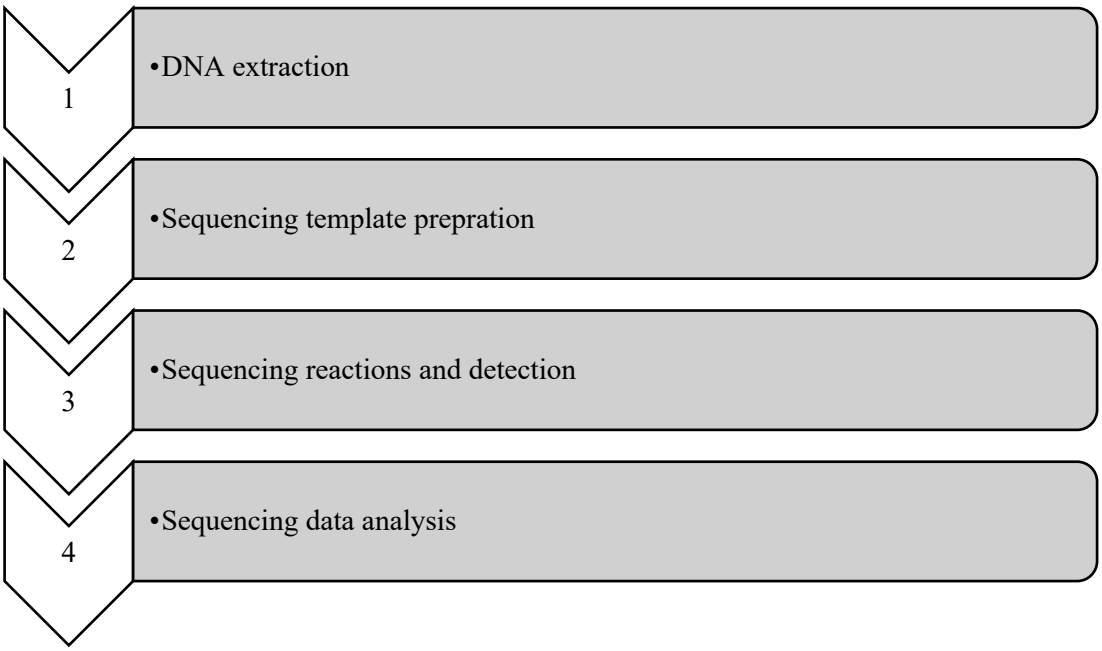



**Figure 2.4:** Workflow of Next Generation Sequencing (NGS).

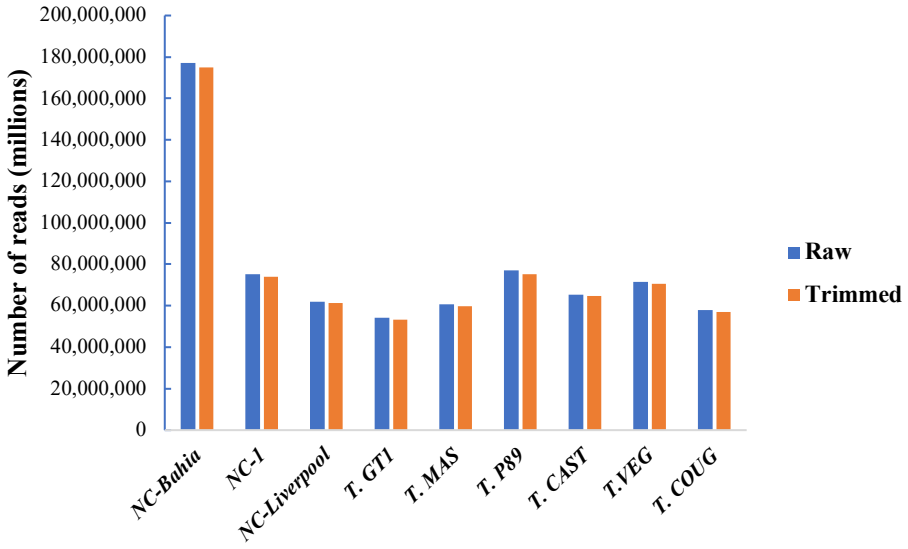



**Figure 2.5**: The total number of reads in millions retrieved from each library of the three *N. caninum* strains (*NC-1, NC-Bahia and NC-Liverpool*) and six *T. gondii* strains (*T. GT1, T. MAS*, *T. P89, T. CAST, T. VEG and T. COUG*). Blue colour indicates to raw reads and orange indicated trimmed reads ready for downstream analysis. The high number of reads in *N. c. Bahia* strain was due to large number of contaminations (further information in chapter 4).

### 2.8.2 Short read alignment to the reference genome sequences

We used Burrows- Wheeler Aligner (http://bio-bwa.sourceforge.net) for short read alignment to map the paired end reads including forward reads (R1) and reverse reads (R2) to the published reference genomes sequences of *N. caninum* strain Liverpool and *T. gondii* strain ME49. Both genomes were downloaded from Toxoplasma databases resources (ToxoDBv29). To generate mapping data, BW algorithm version 0.7.5a –r405 was used to align all the clean reads to the reference genome using the default parameters. The workflow contains several steps, primarily indexing the reference genome, using SAMtools (http://samtools.sourceforge.net) with command fadix to generate the Fasta file index and Picard tools (http://broadinstitute.github.io/picard/) was used to generate sequence dictionary for both genomes.

### 2.8.3 Manipulating the files with SAMtools /Picard/GATK Tools

The unsorted Sequence Alignment Map (SAM) was generated from the mapping output, converted to a sorted SAM file by using Picard tools, with command Sort SAM to generate a coordinate Binary Alignment Map (BAM) file per sample. The BAM files were generated then the duplication was marked in those files. Genome Analysis Toolkit (GATK3) (https://software.broadinstitute.org/gatk/) was used to complete sequence data processing. Two walkers, RealignerTargetCreator then IndelRealinger, were used to treat those BAM files. The target intervals lists were generated from coordinate-sorted and indexed BAM files by using RealignerTargetCreator then a local realignment was performed with the target intervals that were generated previously to transform regions with misalignments into clean data set of reads that were ready for SNP calling approaches. SAMtools -flagstat and QualiMap platform (http://qualimap.bioinfo.cipf.es/archive.html) were used to provide full alignment statistics and additional mapping information to evaluate the alignment data that were sorted in BAM files according to the features of the different types of reads that had passed or failed the quality control (QC).

Each category in the output file was examined more specifically and mapped and unmapped reads for each strain were extracted individually for further analysis. After processing the data, all the sequence coverage files (BAM) were saved then uploaded and viewed in Artemis (http://www.sanger.ac.uk/science/tools/artemis) that was installed locally to visualize the coverage and sequences of the different aligned read sequences for all representative isolates. The coverage and distribution of sequenced reads within different chromosomes was compared to the reference genome that was installed in a local Linux cluster. The schematic overview used for sequencing, mapping strategies and processing of mapped and unmapped reads are summarised in Figures 2.4, 2.5, 2.6 and 2.7.

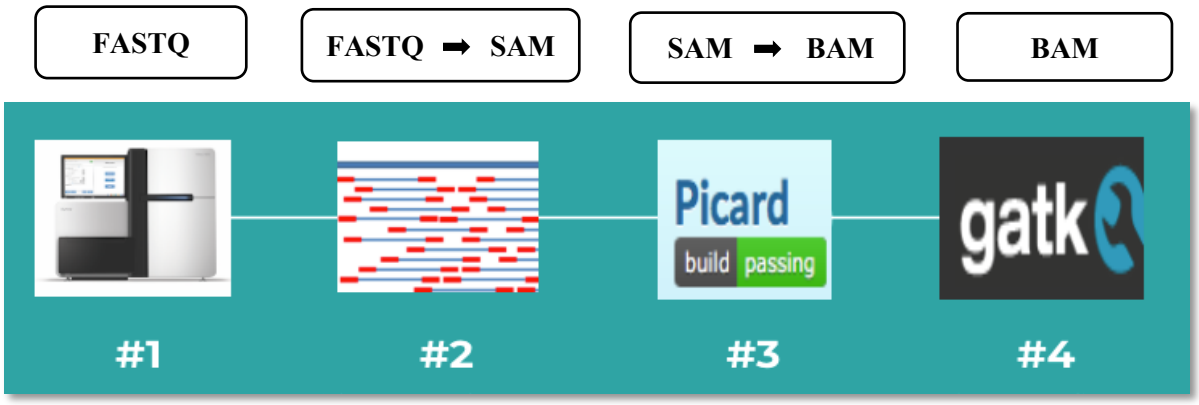



**Figure 2.6:** Overview of typical sequencing data analysis using bioinformatics pipelines that manipulate the files from FASTQ to SAM then to BAM formats through QC and alignment, post alignment, using SAMtools, Picard and GATK tools (http://samtools.sourceforge.net), (http://broadinstitute.github.io/picard/) (https://software.broadinstitute.org/gatk/).

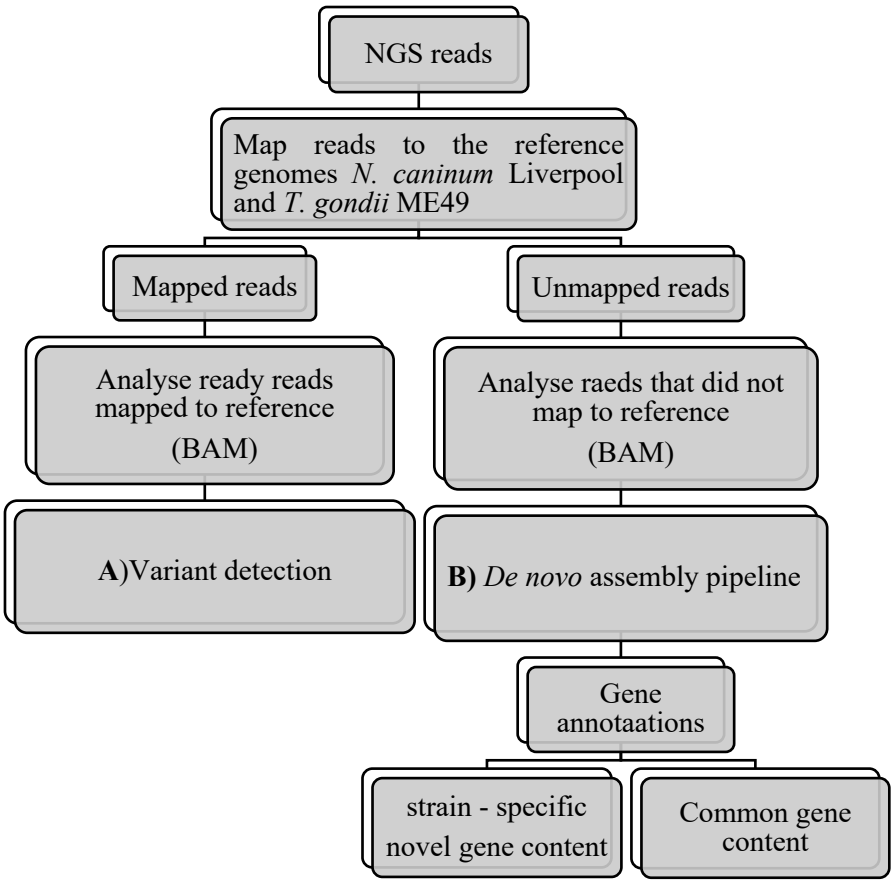



**Figure 2.7:** Comprehensive analysis of mapped and unmapped reads from NGS reads utilized to generate high quality of sequences that will be ready for **A)** calling SNPs and **B)** finding novel genes that were not in the references.

## 2.9 Purified Mapped reads processing

### 2.9.1 SNP discovery per sample

The BAM files generated from the purified mapped reads data were exported to GATK3 software (https://software.broadinstitute.org/gatk/) for identification of single polymorphisms (SNPs). A customised set of commands were used to perform variant discovery workflow. The SNPs were called by running the Haplotype Caller highly recommended from the Genome Analysis Toolkit (GATK3) due to the higher level of accuracy than other caller tools, more specifically in difficult regions for variants in the non-diploid organisms. The raw sets of SNPs were generated and stored in VCF (Variant Call Format) files then filtered by performing Variant Filtration walker to extract all the specific SNPs from each call set per isolate after setting specific parameters. High quality filtered SNPs were identified as those that met criteria for each call set. The final output of SNPs were the SNPs that passed all the filters and were ready for the next downstream analysis. The uniqueness of the final passed SNPs was determined from VCFtools (http://vcftools.sourceforge.net) using VCF-compare and the statistical information was also determined by using further functions from the same tool (VCF-stat) that compared between and within the different strains of *N. caninum* and *T. gondii*. The final stage of the study was of the overlapping and intersection SNPs present in the number of VCF files per sample performed by using BedtoolsV2.26.0 (http://bedtools.readthedocs.io/en/latest/). VCF files were saved then uploaded to be examined by two viewing tools namely Integrative Genomic Viewer (IGV) (http://software.broadinstitute.org/software/igv/) and Artemis (https://www.sanger.ac.uk/science/tools/artemis) to distinguish the real SNPs as a vertical red line compared to the reference genome. The summary of SNP discovery workflow is presented in Figure 2.8.

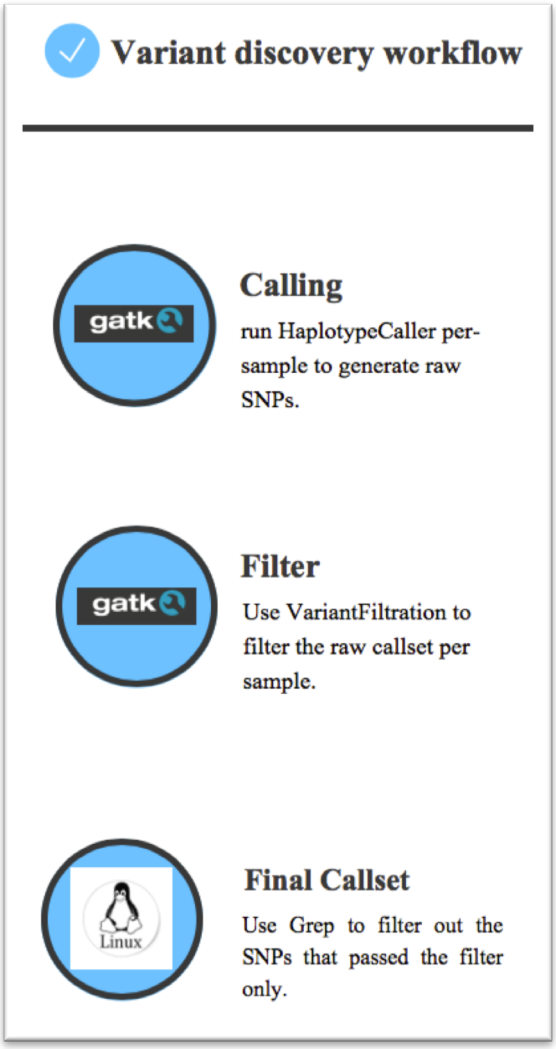



**Figure 2.8**: Summary of SNP detection strategy to generate the final filtered call sets of SNPs that passed all the set of filtrations in all isolates.

### 2.9.2 SNPs functional annotation

Customized Perl scripts and Single- nucleotide polymorphism effect predictor that were run (SnpEff version 4.0 http://snpeff.sourceforge.net/). This program was used to deduce accurate structural and functional annotation of SNPs based on their locations and effects. This program was divided into two stages, firstly to build database then estimate effects. We successfully uploaded the reference genome for both organisms with their annotation files as FASTA and GFF files that were both downloaded from Toxoplasma Genomic Resources (http://toxodb.org/toxo/), to determine the effects on protein coding such as synonymous or non- synonymous mutations or non- coding proteins. All the input files were submitted in VCF format and three different output files with varied format (html, txt, vcf) provided all the potential outputs of SNP information, genetic information and classification of impacts (low, moderate, modifier and high). The low impact they won't change the protein product such as synonymous variants. The moderate impact was those that were unlikely to significantly change the protein produced but might alter interactions with other protein products. Variants labelled modifiers were generally restricted to non-coding regions and have an impact on regulatory regions such as un-translated regions (UTRs). The SNPs with the highest impact were labelled as high impact, and these were presumed to have a greatly disruptive effect on the protein product. All the impacts causing by SNPs whether those effects have a deleterious effect or not and per types for example, (start lost, stop gained) per sample. The abundance of the SNPs per strain was calculated depending on their types and effects by running vcfEffOnePerLine.pl script provided from this software package to divide the VCF files into one line with one effect. The Snpsift (http://snpeff.sourceforge.net/SnpSift.html#intro) programme was applied to filter all the outputs files and plotted individually based on their physical locations (bp) across all 14 chromosomes for all the samples.

## 2.10 Gene ontology enrichment analyses

GO terms were estimated statistically for the gene sets to determine whether specific gene function, cellular contents or biological process were overrepresented or not. The genes containing SNPs grouped into four levels of impacts; low, moderate, modifier and high, were used with ToxoDB's GO term analysis package to assign the GO terms to those genes with a specific impact to understand if particular pathways were more enriched with different sets of impacts in different strains of both species. Only GO terms with a P value of less than 0.05 were assigned in this study. For each set of genes with a particular impact, all the P-value, Bonferroni adjusted P-value and the Benjamini-Hochberg values were calculated per strain.

## 2.11 Fold enrichment (REViGO)

From ToxoDB database, a direct link to Reduce Visualize Gene Ontology (REViGO) software was freely available at (http://revigo.irb.hr/). The GO terms that were identified for all the genes unique to each strain were submitted to this software to reduce the sets of GO terms then visualize to show the specific pathways enriched per strain. The lists of GO categories were tested for statistically significant enrichment per species and the calculation of semantic similarity was measured between the different GO terms that showed the results in different ways. The GO terms that were identified for all the genes having SNPs per impact were also submitted to this software to minimize the sets of GO terms then visualize the specific pathways enriched per strain and per impact. The uniqueness of the data was determined by choosing the degree of allowed medium similarity as 0.7. All the terms that were greater than 0.7 degree were collapsed. However, the clusters with significant pathways were plotted in different colours that reflected the increase in the number of significant enrichments of data. The bubble colour indicates the p-value (blue and green bubbles were GO terms with more significant p-values than the orange and red bubbles). The size of the circle indicate the frequencies of the GO terms and the bubbles of more general terms have the largest size.

## 2.12 Purified unmapped reads processing

### 2.12.1 Identification the composition of unmapped reads

The resulting of non-aligned reads was considered in both *N. caninum* and *T. gondii* strains due to the poor mapping results. To investigate this, the Metagenomic Phylogenetic Analysis (MetaPhlAn) tool was used to identify the composition of the unmapped reads for the samples that proved to have high proportions of unmapped reads that led to falsely mapped reads to the *N. caninum* and *T. gondii* references. The process of detection of the contamination was performed. To identify the most abundant organisms to see whether the top candidates were derived from those reads or not. Due to lack of many eukaryotic species that might ignore the reads entirely and in terms of sensitivity, MetaPhlAn was considered less sensitive than directly aligning the reads. For those reasons, short read alignment (Bowtie2) was applied against the reference genomes of parasite and Mycoplasma bacteria (*M. hyorhinis* species) (https://www.ncbi.nlm.nih.gov/nuccore/CP016817).

To investigate this, the unmapped reads were examined to discover whether the highest proportions of those reads that failed to map was due to missing sequences in the reference genomes, errors in the sequencing methodology or sample contaminations during the experiment in tissue culture. To get more accurate data, the unmapped reads were re-examined by aligning to the references of the two closely parasites (*N. caninum* and *T. gondii*) and a bacterial genome *M. hyorhinis.* Unmapped reads still remained that were consider ambiguous taxonomic sequences. Blast searchs were therefore performed against the (nr) database using the default parameters. Reads were identified belonging to the host genome (Vero Cell line). This occurred when we increased the sensitivity of BLAST parameters. Following that, we aligned them to the reference of green monkey (*Chlorocebus sabeus*) genome that was downloaded from https://www.ensembl.org/Chlorocebus_sabaeus/Info/Annotation, which is the host genome for Vero cell line. To produce a more sensitive overview of the reads that were still not mapped, a further alignment step was applied using Bowtie2 again to align all the reads from each sample against the combined parasite *M. hyorhinis* and host to derive clean set of reads that did not belong to the parasite or bacteria and host. Re-examining the non -parasite and non-bacteria sequences with relaxed parameters (low complexity filter switched off and scores modification) were performed.

All the host contamination was removed from the reads to produce a comprehensive set of reads without host contamination that was extracted with SAMtools 0.1.18, then saved in a separate output file for filtering them and then de novo assembly using alter *De novo*-based pipeline.

### 2.12.2 *De novo* assembly of unmapped reads pipeline

Due to the small size of the genome of both parasites, ability of handling with Illumina paired-end-sequences, memory and time intensive were considered. SPAdes Version 3.11.1 (http://spades.bioinf.spbau.ru/release3.11.1/manual.html) was used that was highly effective as a *De novo* genomic assembly tool. This approach was locally installed on a core Linux cluster. The workflow of this pipeline was primarily relied on to convert the reads to k-mers to start de Bruijn graph and then assembling them into error-correcting contigs and scaffolds. The accepted Illumina reads were entered into a *De novo* assembly as an input using default settings as the different stages were performed. All SPAdes outputs were stored to determine the quality of assemblies produced.

### 2.12.3 Assembly evaluation of the final outputs

For the assembly output files were generated, Quality Assessment Tool for Genome Assemblies (QUAST) (http://quast.bioinf.spbau.ru) was used to evaluate the quality of the assemblies produced. By looking at the summary statistics that had been generated using the scaffolds and contigs files as an input, the outputs of these assessments were compared for different metrics such as; N50, NG50, the maximum length of the contigs and GC% to assess the quality of the assembly.

### 2.12.4 Integration of Genome Assemblies (Blob-tools)

To distinguish if there was still a large quantity of contamination in assembly. The Integration of Genome Assemblies (Blob-tools) (https://github.com/DRL/blobtools) was performed by running a series of commands with settings by the user to identify parasite reads without contaminates reads. This should a comprehensive representation of the target genome derived from the parasite sequences only by checking and filtering all the assembly reads based on the GC content, coverage depth and taxonomic annotations. Briefly, the *De novo* protocol was used in this study based on user input files that had constructed the scaffold and contigs into BlobDB data, then plotted them in different coloured dots that were ranked by taxonomic BLAST of the profile contig output. The coloured circles were positioned in two axes; the X-axis based on their GC proportion and the Y-axis showed the sum of read coverage of the contigs. To improve the final output assembly set, many rounds were performed to reduce the set of reads that did not belong to the target sequences and to get the final high-quality assembly output without contamination that was most likely to be derived from the *N. caninum* and *T. gondii* genomes only.

## 2.13 Gene finding and annotations pipeline

To annotate the genes that were found in the scaffold and contigs files after assembling the reads, we used the Companion gene annotation tool (http://companion. sanger.ac.uk) to generate and visualize new features of annotation for the *N. caninum* and *T. gondii* genomes. The input files were prepared and entered in FASTA format as target sequences. The criteria for quality of sequence files were considered against the two published reference genomes that were available in this tool as target references which were originally imported from the public databases GeneDB and EuPathDB. An extensive results file in different format was downloaded and saved for downstream analyses. The Companion pipeline performed an overview of plots that demonstrated a wide range of the reference-target alignments including coding and non-coding genes, gaps and some unique genes that were not found in the target genomes or missing core genes that were distributed across the 14 chromosomes through rearrangement. BLAST searches were performed to all the sequences identified with parameters settings against (nr) databases to find if there were any

significant hits to indicate identical sequences between the sequences entered and the sequences in the database. All the sequences were evaluated to identify the level of similarity and overlap per gene found in each strain. Phylogeny trees of the unknown sequences were displayed to determine the distances between the groups of sequence homologs or not. According to the distance from the varied queried sequences in the output of the BLAST search, the unknown genes were considered as strain-specific sequences or common contigs based on the identity percentage that was identified in all the list of genes per strain.

### 2.14 Copy Number of Variations estimation

We used CNVantor for detection of copy number variation (CNV) that was suitable for the paired-end data from the Illumina platform based on read depth (RD). Thus, tool was freely available at (http:// sv.gersteinlab.org/cnvnator) and applied to *N. caninum* and *T. gondii* genomes that were parsed from BAM files. A series of commands using read depth values from the BAM files were entered as an input file. All the statistics and plots were generated and examined manually to identify all the genes that were located in specific regions of the genomes to calculate the overall estimations of duplication and deletion events across the genomes per strain. One of the advantages of this tool is to determine the bin size which is the regions of duplicated or deleted regions for the read depth specified from small to large basses. This tool also suitable for the NGS data and platforms (see section 1.14).

# Chapter 3: Comparative genomic analysis of *T. gondii* and *N. caninum* parasites.

## 3.1 Introduction

The development of the next-generation sequencing technologies, also known as high-throughput sequencing, has dramatically expanded our ability to carry out genomic research in infection diseases and enable functional and clinical studies. The major advantages of NGS technologies are that the methods can detect all the changes in the entire genome, including mutations, duplications and deletion in a cheap, fast and accurate manner (Anderson and Schrijver, 2010; Nowrousian, 2010; Chen *et al*., 2013; Buermans and Den Dunnen, 2014; Nevado, Ramos-Onsins and Perez-Enciso, 2014; Ari and Arikan, 2016; Singh, 2017; Thankaswamy-Kosalai, Sen and Nookaew, 2017). As mentioned in Chapter 1 (section 1.8.3), several comprehensive comparative genomic studies have been performed for parasites that emphasise how comparative genomic and functional analysis based on the NGS approaches have led to significant advances in our understanding of the biology of the parasites. Recent investigations was done by Ramaprasad *et al*., (2015) using RNA-seq datasets that improved the genome annotations of the tachyzoites of *N. caninum* strain *LIV* and *T. gondii* strain *VEG* parasites that leads to accurate gene models (Wasmuth *et al*., 2009; Behnke *et al*., 2011; Reid *et al*., 2012; Adomako-Ankomah *et al*., 2014; Ramaprasad *et al*., 2015).

This chapter describes an effort to further identify genes and genes family members in *T. gondii* and *N. caninum*. It was important to carry out bioinformatic analyses to generate a comprehensive comparative genomic analysis of both species by performing an update of predicted protein lists for both species that have been obtained continuous of the genome annotation as part of a collaborative effort adding to ToxoDB database. There was an opportunity to exploit this information and highlight orthologous relationships, species-specific families and unique genes that had not already been identified in the previous annotations or in the published literature and also providing corrections to the existing annotations for both species.

Such new findings generated from our comparative genomic analysis might offer greater insight into the virulence associated factors involved in host-parasite interactions. Several unanswered questions remain about the mechanism of infection, geographical distributions, transmission, pathogenesis and the number of hosts. One of the important questions is why *T. gondii* has a wide range of hosts but *N. caninum* does not, and what are the genetic differences between them. (El Hajj *et al*., 2007; Reid *et al*., 2012; Kemp, Yamamoto and Soldati-Favre, 2013; Lei *et al*., 2014; Reid, 2015; Shwab *et al*., 2016).

### 3.2 Aims of this Chapter

This chapter describes an effort to identify orthologues, species-specific and unique genes based on a comparison between the published reference genomes. In this chapter, I will use the *T. gondii* genome strain ME49 and *N. caninum* strain Liverpool in a comparative genomic analysis for both genomes in order to investigate the changes that might have an impact on the biology of the parasites and to give a greater understanding of the divergence between them and the genomic features that might underlie phenomena such as virulence, host restriction, host - parasite interaction, transmission and disease manifestations. An initial comparative genomic analysis of the two species was published in 2012 (Reid *et al*., 2012). Since then new versions of the *Toxoplasma* genome have been released via the *Toxoplasma* genomic resource (ToxoDBv29). Hence the three aims of this chapter are:

1. To revisit the comparative genomic analysis between *T. gondii* and *N. caninum*.
2. To identify the shared gene set between the two species*; T. gondii* and *N. caninum*.
3. To identify species-specific genes unique to each species, which might play a direct role in divergence by highlighting insight into expansion of the distinct gene families that are often associated with pathogenesis, adaptation, host preference and parasite survival styles.

## 3.3 Results

### 3.3.1 Identifying orthologue cluster between *T. gondii* and *N. caninum*

In order to revisit the comparative genomic analysis, we ran OrthoMCL with default parameters between the two protein sets that were obtained from ToxoDBv29. Given that the aim of the task was to identify orthologous, species - specific and unique (unclustered) genes, the final OrthoMCL output was generated using the default MCL values for clustering as explained in Chapter 2 (section 2.1.1). The results are summarised in Figure 3.1. The total number of sequences included in the analyses was 15,447 from the two genomes consisting of 8,322 and 7,125 protein-coding genes in each *T. gondii* and *N. caninum*, respectively. A total of 13,600 sequences were clustered into 6,518 orthologous groups with 1,847 sequences remaining unclustered. A total of 13,463 orthologous genes were shared between the two species. The largest group within those clusters was the group which had one - to - one orthologous relationship containing 12,678 genes (94%) in 6,339 clusters. Our results hence indicated that the gene content was largely conserved between *T. gondii* and *N. caninum*. In both genomes, the vast majority of genes were annotated as 'hypothetical' or 'conserved hypothetical' proteins.

When expanding the protein comparison between the two species to include further orthologous genes then this second group contained 142 clusters comprised of 785 (6%) genes between the two taxa. The analysis of those shared multiple - gene families between the species within the core gene sets revealed several interesting features. Comparison of OrthoMCL groupings identified a significant increase in members of multi-gene families reflecting the expansion of surface antigen protein encoding genes in *N. caninum*. Most protein-coding genes in *T. gondii* belonged to the *Toxoplasma gondii* family proteins, rhoptry proteins, dense granules and surface antigen proteins genes. The remaining 137 sequences comprised 37 clusters that contained species - specific gene families in *T. gondii* and *N. caninum* with more than one copy including 98 and 39 genes that were unique to *T. gondii* and *N. caninum*, respectively. In *T. gondii*, more than half of the 98 genes encode proteins of unknown function (53%), the most prevalent families encodes the *Toxoplasma gondii* family proteins with a total of 26 TgFAMs genes (26.5%) containing three subfamilies; TgFAM C (14 genes);

TgFAM A (10 genes) and two genes belonging to TgFAM B. The remaining genes belonged to the KRUF (9%), SRS (4%) and GRA (2%) gene families. No MIC genes were noticed in this group of unique genes. In *N. caninum*, the greatest number of unique genes were encoding surface antigen proteins, totalling 25 (64%) genes. The second largest group of genes encoded proteins of unknown function (20%). A total of 1,847 genes were found to be true singletons, ie single - copy genes that were specific to either *T. gondii* or *N. caninum*; 1,536 (83%) were singletons in *T. gondii* and 311 (17%) in *N. caninum* without any paralogues or orthologous based on our orthologous analysis. Table 3.1 and Figures 3.2 and 3.3 compares the number of unique genes identified in this project with Reid *et al*., (2012).

Reid *et al*., (2012) reported a total of 7121 and 7286 protein coding genes in *N. caninum* and *T. gondii*, respectively. Using manual curation, the list of species-specific genes was narrowed down to 113 in *N. caninum* and 231 in *T. gondii,* the majority of which encoded proteins of unknown function. Out of those, 72 and 43 predicted proteins had Pfam domains. A further subset (43 out of 113 in *N. caninum*) had additional proteomic - based evidence to indicate expression.

The comparison of the species - specific genes as extracted from Reid's analyses and our gene set highlighted some discrepancies (see also Table 3.1). We identified a large number of species - specific genes, 291 in the *N. caninum* and 1,544 in the *T. gondii* genomes. In *T. gondii*, we found an increase of 12.4% in the number of protein coding genes compared to those previously reported by Reid *et al*.,(2012). We also found a significant increase in the number of protein coding genes in *N. caninum* compared to the Reid analysis (Figure 3.2).

From our findings, a total of 313 unique genes were identified after removing the surface antigen family. A total of 22 genes were shared between the Reid set and our set of unique genes. A larger number of unique genes were found in our analysis, reflecting an additional number of species -specific genes added to the *N. caninum* genome, most encoding 'hypothetical' and 'conserved hypothetical' proteins of unknown function.

In *T. gondii*, we identified a higher number of unique genes than in the Reid dataset. A similar approach has been applied by removing the SRS genes from our list. In Reid list for *T. gondii* there was 231 genes, of which a total of 59 were shared with our set of 1,544 genes (Figure 3.3). Notably, a large number of *Toxoplasma gondii* family proteins was noticed in our species-specific gene list consistent with the notion of diversification of gene families. This family was previously called the as *Toxoplasma* -specific-family (TSF) in Reid's study with a lesser number of members (see also below). In addition to those gene families, other gene families noticed in our list includes ROPs (ROP4, ROP16, ROP18 and ROP39), GRAs (GRA11 and GRA12) and 10 members of the KRUF protein family ( see dataset of chapter three available in appendices; (A) include: summary of the OrthoMCL clustering of *T. gondii* and *N. caninum*.

**Table 3.1:** The comparison of protein coding and species-specific genes from Reid *et al*., (2012) and our analysis of the two species *T. gondii* (*ME49*) and *N. caninum* (*Liverpool*).

| | **Reid *et al.,* (2012)** | | **Our analysis (ToxoDBv29)** | |
|---|---|---|---|---|
| **Genes** | ***T. gondii*** | ***N. caninum*** | ***T. gondii*** | ***N. caninum*** |
| **Protein coding** | 7,286 | 7,121 | 8,322 | 7,125 |
| **Unique (species-specific)** | 231 | 113 | 1,544 | 291 |

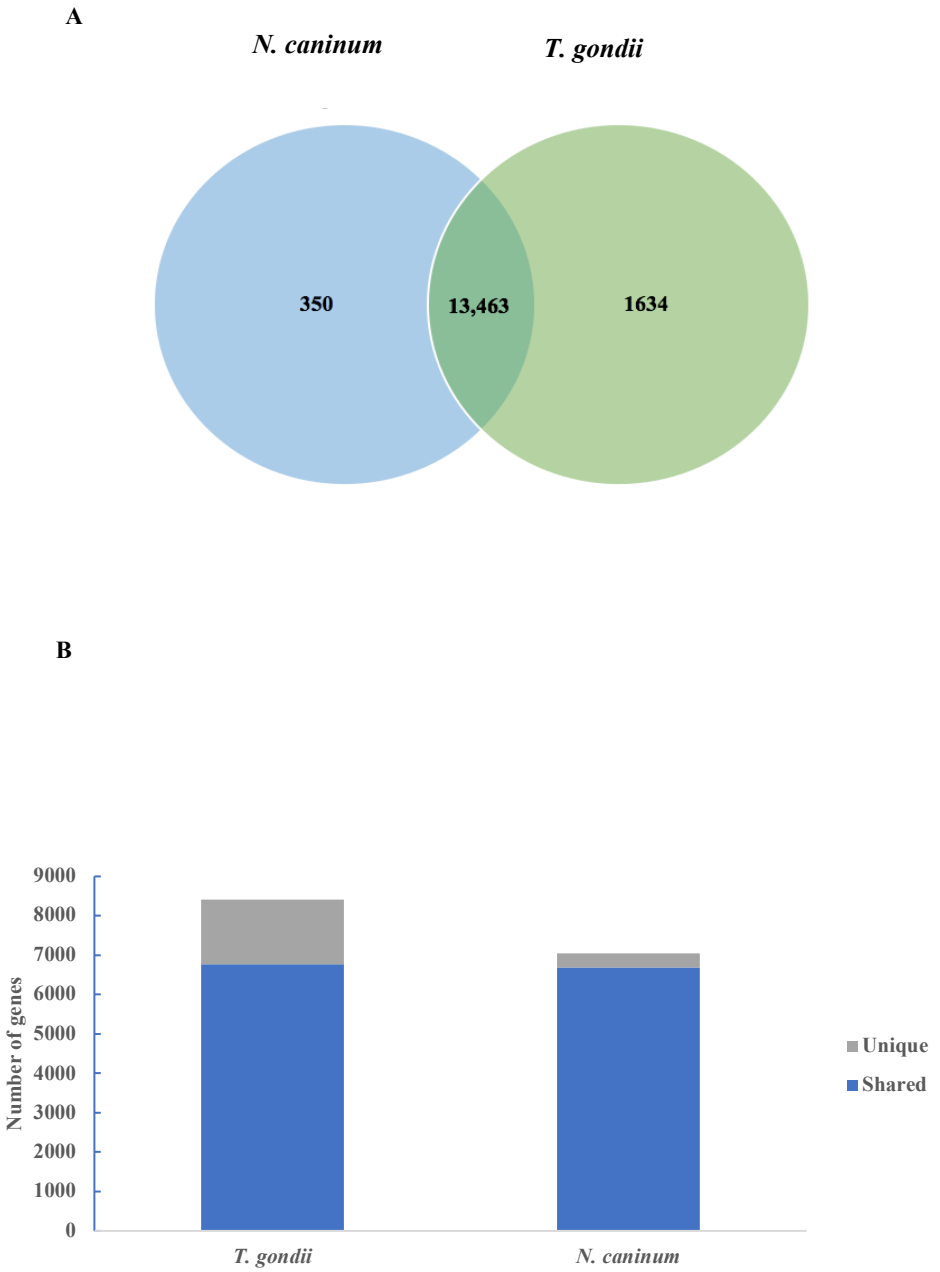



**Figure 3.1: A)** Number of shared and unique genes as defined by OrthoMCL in *T. gondii* and *N. caninum*. The shared13,463 genes include 12,678 genes (one-to one orthologues) and 785 genes shared more than one copy **B)** Analysis of shared and unique genes as predicted by OrthoMCL in *T. gondii* and *N. caninum*. Blue bars represented orthologous genes and identifies those shared between the two species. Grey bars indicate unique genes with no paralogues or orthologues per species.

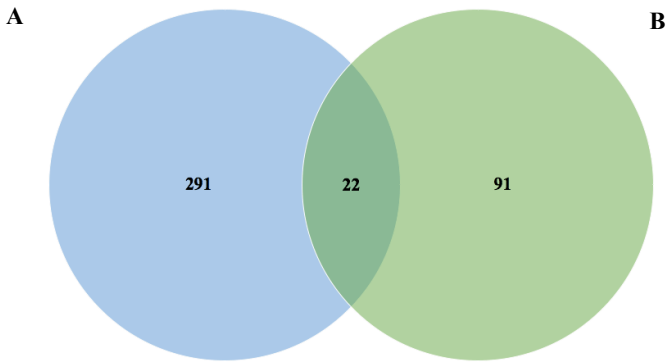



**Figure 3.2:** The number of unique genes in *N. caninum*. **A)** Blue circle indicates the number of unique genes from our analysis; **B)** Green circle indicates to the number of unique genes from Reid et al., 2012. A total of 22 genes were shared between the two sets of unique genes.

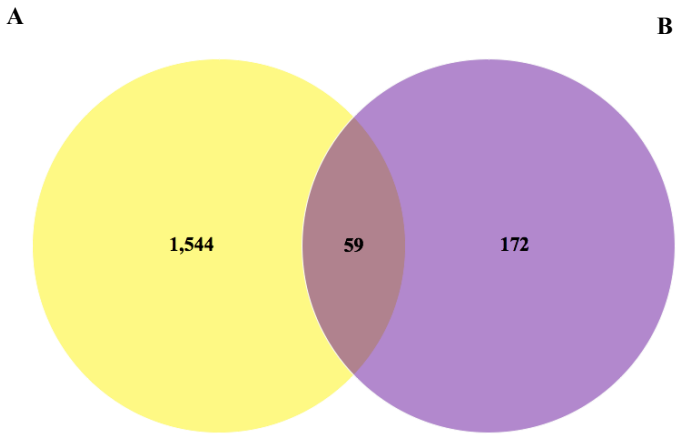



**Figure 3.3:** The number of shared and unique genes in *T. gondii*. **A)** Yellow circle indicates the number of unique genes from our analysis; **B)** Purble circle indicates the number of unique genes from Reid et al., 2012. A total of 59 genes were shared between the two sets of unique genes.

### 3.3.2 Identification of key biologically relevant gene families in *T. gondii* and *N. caninum*

#### 3.3.2.1 Identification of Surface antigen proteins (SRS)

We manually checked and compared the total number of the superfamily of glycosylphosphatidylinositol (GPI)-anchored proteins known as SAG1-related sequences (SRS) genes that were significantly expanded in both parasites with those members that were reported previously (Reid *et al*., 2012). Having identified the recent number of SRS genes belonging to *T. gondii* and *N. caninum* species respectively, we carried out a comparisons between the SRS genes from Reid *et al*., (2012) and ToxoDBv29 (our lists) to identify reductions or expansions between those protein families.

In the Reid analysis, a total of 104 and 227 SRS genes were reported that were distributed among 14 chromosomes in *T. gondii* and *N. caninum,* respectively. We identified additional members of the SRS family. This indicated that improvements have been made to the predicted *T. gondii* and *N. caninum* genes models for both parasites. In addition to the previously recognized SRS genes, we found a further 7 and 10 SRS genes in *T. gondii* and *N. caninum*, respectively (Table 3.2). Differences in these families might play a significant role in host range restriction and pathogenesis as previously postulated by Reid *et al*., (2012). An expansion of this gene family in *N. caninum* has also been associated with the host restrictions, metabolic pathways and pathogenesis (Howe *et al*., 1998; Reid *et al*., 2012; Reid, 2015).

Not surprisingly, it has been found that these genes are involved in antigenic variation, which is potentially responsible for genetic diversity, life style, adaptation and host parasite interactions and also in switching aspects of parasite and host life (Risco-Castillo *et al*., 2011; Wasmuth *et al*., 2012). It has been noticed that large repertoires of this gene family were often localized in telomeric regions, suggesting that there are positive associations between this pathogenic gene family and telomeres (Barry *et al*., 2003; Forrester and Hall, 2014).

**Table 3.2:** The comparison of the surface antigen families (SRS) genes in both species between the two studies.

| Species | Reid *et al*., 2012 | ToxoDBv29 |
|---|---|---|
| ***T. gondii*** | **104** | **111** |
| ***N. caninum*** | **227** | **237** |

### 3.3.2.2 Identification of Rhoptry gene family (ROPs)

In addition to the SRS gene family, there were further differences in another virulence gene family known as rhoptry kinase genes. As we mention before in Chapter 1, this group of proteins is involved in host - pathogen interaction by discharging the contents released from the rhoptries organelles. The comparison of the putative ROP genes provided from Reid's work and our ROP genes highlighted fewer ROP genes in our analysis (Table 3.3). We noticed that there was a subset of genes that did not have a syntenic orthologue with in the other species but were rather exclusively present in one species. After comparing the list of putative ROP genes in *T. gondii* strain ME49 to Reid's list, a total of 64 ROP genes were identified, 11 of which were annotated as rhoptry neck proteins. In *N. caninum*, a total of 58 ROP genes were observed from our putative list, however 68 genes were reported by Reid *et al.,* (2012).

It is worth highlighting that there were differences in the number of specific ROP genes that were found in one species only, specifically, we found that there was a group of species - specific genes in both organisms as confirmed earlier from OrthoMCL. From Reid's list, ROP2A, ROP2B, BRP1, ROP8, ROP18, ROP42 and ROP43 were considered *Toxoplasma* - specific genes. Our list of ROP genes revealed fewer species specific ROP protein than was reported earlier. We believe this discrepancy can be attribute to incorrect annotation in the previous reference genome release in Reid's annotations with a group of ROP42, ROP43 and ROP44 having been incorrectly annotated as ROP8.

In *N. caninum*, a group of genes specific to this species includes ROP1B, ROP5B, ROP15B, ROP51, ROP52, ROP53 and ROP55. However, from our analysis it seems that no ROP8, ROP2A, ROP2B and ROP18 genes were present in the *N. caninum* genome. Notably, ROB5B, which is paralogous to ROP5 was incorrectly annotated as ROP2-related. ROP51 was indeed a species-specific gene for *N. caninum* as also reported by Reid *et al.* (2012). Furthermore, we found that there were additional ROP proteins not presented in the Reid list but considered as *Toxoplasma* - specific proteins in our list, including ROP39 (TGME49_062050). Two ROP members (TGME49_09602) and (TGME49_096000) were also observed in the *T. gondii* genome but not in *N. caninum*.

Lastly, three members of the ROP gene family were observed in both of the two lists; this is a significant set of genes as the encoded proteins play an important role in the virulence variations between the two coccidian parasites (Reid *et al.*, 2012). From both lists it was clear that ROP5 and ROP16 were annotated in both genome release. However, one specific ROP gene (ROP18) considered as a major virulence factor present in the genome of *T. gondii* as a functional copy but truncated by a premature stop codon in the *N. caninum* genome; our analysis confirmed Reid's initial findings, which considered this gene a *T. gondii*-specific gene.

Further comparison between the two candidate virulence genes revealed that there was only one gene in *T. gondii* annotated as ROP5 (TGME49_308090) in our list. In Reid's list, no ROP5B gene was identified in *T. gondii*. However, in *N. caninum*, we noticed that there were multi-gene copies of ROP5, named ROP5A and ROP5B (NCLIV_060730) and (NCLIV_060740). Subsequent chapters will describe how this ROP5 was found to be tandemly repeated in different strains of *T. gondii.* In *N. caninum,* one copy of ROP 5 was annotated (NCLIV_060741). ROP 16 was noticed in both lists for both species annotated as TGME49_262730 an orthologue of NcROP16 (NCLIV_025120) (Ong, Reese and Boothroyd, 2010; Ma *et al.*, 2017b).

**Table 3.3:** The comparison of the rhoptry genes families (ROPs) genes in both species between the two studies.

| Species | Reid *et al*., 2012 | ToxoDBv29 |
|---|---|---|
| ***T. gondii*** | **68** | **64** |
| ***N. caninum*** | **68** | **58** |

#### 3.3.2.3 Identification of *Toxoplasma gondii* family protein (TgFAMs)

The Reid study described the *Toxoplasma*-specific family that is comprised of ten members (TgTSF1 - TgTSF10) that were largely absent from the *N. caninum* genome. However, we identified that there were four further TSF members that had not been reported in the Reid list and named as *Toxoplasma gondii* families proteins (TgFAMs) group C (TgFAMAC) in subsequent releases of the *Toxoplasma* genome with a total of 14 genes that were absent from *N. caninum* genome. As we expected, changes in the gene annotations had an effect on the number of the TgFAMs genes identified. Our findings highlighted a significant expansion in this specific family with four additional subsets of *T. gondii* gene families includes TgFAMAs, TgFAMBs, TgFAMDs and TgFAMEs, totalling 81 genes distributed across the 14 chromosomes and in the unassigned contigs (Table 3.4). Interestingly, a high proportion of the TgFAMAs genes were tandemly duplicated and clustered in two specific chromosomes; XII and VI with more than half out of the total genes identified.

Analysis of orthologous group clustering was performed to test the level of homology among those two species. By further examining the orthologous groups, we noticed that there was a high level of synteny between these genes among *T. gondii* and *N. caninum*, specifically in groups A, B, D and E. However, all the 14 members belonging to the TgFAMC family genes were considered *T. gondii* unique family genes. In line with Reid's list and recent investigations done by Lorenzi *et al.,* 2016, our analysis presented further evidence of the presence of a subset of TgFAMs known now as TgFAMC that was exclusively located in the *T. gondii* genome which led us to distinguish *T. gondii* from the *N. caninum* species.

**Table 3.4:** The comparison of the *Toxoplasma gondii* families (TgFAMs) genes in both species between the two studies. In Reid *et al*., 2012**,** this family was known as *Toxoplasma* specific family (TSF).

| **Species** | **Reid *et al*., 2012** | **ToxoDBv29** |
|---|---|---|
| ***T. gondii*** | **10** | **81** |
| ***N. caninum*** | **7** | **67** |

### 3.3.2.4 Identification of Dense granule genes family (GRAs)

In Reid's putative GRA list, the total number of Dense granules proteins (GRA) was 17 and 15 GRA genes in *T. gondii* and *N. caninum*, respectively. We identified a total of 18 and 20 GRA genes (Table 3.5). Both lists confirmed the absence of the two GRA genes GRA11 and GRA12 from *N. caninum*, this confirming that both genes were considered *T. gondii* - specific genes. However, our data showed that there was one further GRA11, known as GRA11B (TGME49_237800). In this case, we can named them as GRA11A (TGME49_212140) and GRA11B (TGME49_237800), in line with a recent analysis (Ramakrishnan *et al*., 2017). In addition, Reid *et al*., (2012) consider GRA12 a single gene with no orthologues, as we mention above, but, two GRA12 genes (TGME49_278850 and TGME49_288650) were identified in our analyses.

Interestingly, one GRA12 (TGME49_288650) had was orthologous to NCLIV_041120 that was annotated as a conserved hypothetical protein in *N. caninum*, information not reported in the Reid list. In addition to this, no syntenic relationship was observed with the additional GRA12 gene (TGME49_278850) (Kolben, Maurer and Knitza, 2004; Michelin *et al*., 2009). No GRA13 was found in both species. Further GRA genes that were also not in the Reid list were two members of the DG32 gene family (TGME49_22217 and TGME49_297880) and GRA15 which was only identified in *T. gondii*. In *N. caninum*, one further member was not in Reid list, a gene encoding an NTPase protein.

**Table 3.5:** The comparison of the Dense granule genes families (GRA) in both species between the two studies.

| Species | Reid *et al*., 2012 | ToxoDBv29 |
|---|---|---|
| ***T. gondii*** | **17** | **18** |
| ***N. caninum*** | **15** | **16** |

### 3.3.2.5 Identification of Micronemes genes (MICs)

Further comparisons of the putative MICs gene list were performed. A total of 33 and 38 MICs genes were provided from Reid's list for *T. gondii* and *N. caninum*, respectively with the significant absence of five MICs members from *T. gondii* namely MIC26, MIC19, MCP5, MCP6 and MCP7, which meant those genes were critically considered as *N. caninum* specific genes at that time. In the current study a total of 39 and 42 MIC genes were identified in *T. gondii* and *N. caninum* respectively, that overlapped with the Reid list. We found no significant evidence of specific microneme genes in *T. gondii* identified from our list (Table 3.6).

However, in *N. caninum* from both studies, a group of previously reported micronemes putative proteins MIC15, MIC16, MIC20, MIC21, MIC22, MIC24 and MIC25 genes were observed in our current list. This suggested the identification of this *N. caninum* - specific groups was similar in both studies. An additional group noticed as absent in Reid's list including MCPs genes (MCP5, MCP6 and MCP7) *T. gondii* and also from our list of *T. gondii*.

More significantly, it was important to point out that the MIC14 protein was absent from *T. gondii* based on Reid *et al*. (2012) findings. However, our annotation indicated that that microneme protein MIC2 (NCLIV_022970) had a paralogue known as MIC2B (NCLIV_033690) which annotated as MIC14 in Reid *et al* but is currently annotated as MIC26. The MIC2B gene was considered a *N. caninum* specific-gene with no orthologues in *T. gondii*. In *T. gondii*, it has been confirmed that MIC2 (TGME49_201780) is involved in gliding motility and invasion by working with a MIC2- associated gene that is known as M2AP (Tonkin *et al*., 2010; Huynh *et al*., 2015).

**Table 3.6:** The comparison of the Micronemes gene families (MIC) genes in both species between the two studies.

| Species | Reid *et al*., 2012 | ToxoDBv29 |
|---|---|---|
| ***T. gondii*** | **33** | **39** |
| ***N. caninum*** | **38** | **42** |

#### 3.3.2.6 Identification of other gene families between species

Given the findings described above, other protein gene families were identified in both putative gene lists that have a much smaller number of members than the SRS, ROP, TgFAM, GRA and MIC gene families. In Reid *et al* (2012) findings, it was reported that there was a gene family named SAG-Unrelated Surface Antigen (SUSA) that was involved in antigenic variation (Pollard *et al*., 2008). The number of annotated genes per genome at that time was 26 and 38 in *T. gondii* and *N. caninum*, respectively. Pollard *et al*., (2008) mentioned this family of glycosylphosphatidylinositol-anchored surface antigens with that clustered on chromosomes VI and XII. We now postulate that the dramatic increase of the TgFAMs gene family mentioned earlier included SUSA genes that were recently re-named to TgFAM, more specifically TgFAMA, which includes a total of 33 genes. Two SUSA genes that were highlighted in Reid's list (NCLIV_067570) and (NCLIV_067920) both have a 1:1 orthologue with genes in *Toxoplasma* (TGME49_2780800).

In addition to this family, Reid mentioned in his findings that there was a novel family named Lysine-Arginine rich Unidentified Function (KRUF), which had three members in *N. caninum* and that was expanded to 7 genes in *T. gondii*. Our data showed that there was an identical number of genes in *N. caninum*. Double the number of KRUF gene were noticed in *T. gondii,* suggesting that there was a significant increase in the number of KRUF in *T. gondii*.

## 3.4 Discussion

Our work has highlighted the importance of revisiting comparative genomic analyses in light of new versions of genome assemblies being publicly released. Our results provide novel insights into the similarities and differences between previously published work by Reid *et al*., (2012) and our own analyses. It is important to review the level of similarity between these two apicomplexan species to determine the presence of shared gene content and the small differences in the number of unique genes as this will enable us to pinpoint the genomic variations that underlie phenotypic characteristics such as the mechanisms of pathogenesis, host restriction and mode of transmission.

In accordance with the current results, previous investigations have demonstrated that the differences between the two organisms was due to the large expansion of species - specific gene families, often clustered in telomeric regions, having a key role in pathogenesis (Barry *et al*., 2003; Paterson, Vogwill, Buckling, Benmayor, Andrew J. Spiers, *et al*., 2010; Debarry and Kissinger, 2011; Reid *et al*., 2012; Su *et al*., 2012; DeBarry and Kissinger, 2014; Forrester and Hall, 2014; Lorenzi *et al*., 2016). Our data clearly show that – as annotations change and are updated over time - more species-specific genes in *T. gondii* and *N. caninum* can be identified. We were also able to refine functional annotation assignments and correct erroneous annotations carried out by different research communities that can lead to conflicting or differing results in the number of species-specific gene sets when querying the genome databases. Given the improvement seen in the current gene annotations based on manual annotation and review, it is likely to expect that experimental data would further improve the final structural and functional annotations and help to determine the true number of species-specific genes encoded by the *T. gondii* and *N. caninum* genomes.

These results further support the idea of the close relationship between *T. gondii* and *N. caninum* due to the large number of conserved regions and reflect the fact that there is high synteny between the two genomes with one to one orthologue. Comparison of the genes identified in this project with those of previous studies confirmed that the significant differences between the two species appeared to be attributable to the expansion of unique sets of genes, significantly in the number of members belonging

to gene families that include SRSs, ROPs, MICs ,GRAs ,TgFAMs and KRUFs. Such gene families give each genome a specific architecture that may be involved in regulation of virulence, host parasites interactions, genetic diversity, host preference and gene expression (Debarry and Kissinger, 2011; Reid *et al*., 2012; Reid, 2015; Lorenzi *et al*., 2016).

Particularly noteworthy is the large expansion of the SRS gene family in *N. caninum* with more members than reported earlier by Reid *et al*. (2012). In the case of *T. gondii*, the large number of SRS genes are thought to enable the infection of a wider range of hosts (Reid *et al*., 2012). Our data found that the number of SRS genes was significantly higher in *N. caninum* and led us to support the hypothesis of the restriction in the host range of *N. caninum* being related to this large expansion of the SRS gene family. This might answer the question as to why *N. caninum* has a limited number of hosts compared to *T. gondii*. More importantly, the variations (SPNs and CNVs) within this gene family were investigated to assess whether accumulation of variations in the genomes of *T. gondii* and *N. caninum* contribute to this phenomenon (see Chapters 4 and 5).

Lastly, what is surprising is that a newly named gene family was highly expanded in the genome of the *T. gondii* with five sub sets of genes family known as TgFAMs (A, B, C, D and E). As detailed, we postulate that these genes were previously named the TSF and SUSA family with fewer genes predicted. More significantly, TgFAMC family member genes were only noticed in *T. gondii* without any orthologues in *N. caninum*. This finding was in agreement with the results done by Lorenzi *et al*. (2016) who found that the TgFAMs family was considered a unique, tandemly clustered gene family, predominantly clustered on chromosomes XII and VI in the *T. gondii* genome. This family might be involved in adaptations, mechanism of immunity and transmission during sexual development in the definitive host (Behnke *et al*., 2014; Dalmasso *et al*., 2014; Lorenzi *et al*., 2016).

# Chapter 4: Use of multiple strain sequencing to define variants contributing to phenotyping changes among *N. caninum* isolates.

## 4.1 Introduction

Recently studies have increasingly used single nucleotide polymorphism markers to get more understand of the population structure of *N. caninum* strains compared the among isolates worldwide (Regidor-Cerrillo *et al.*, 2006, 2013; Al-Qassab *et al.*, 2009; Al-Qassab, Reichel and Ellis, 2010; Salehi, Gottstein and Haddadzadeh, 2015; Calarco, Barratt and Ellis, 2018). This approach has been enable by advances in sequencing technologies and offers advantages over microsatellites (Short tandem repeats or STRs). These strategies allow for the identification of thousands to millions of unbiased mutations, and the simultaneous estimation of SNP frequencies across the genomes in different chromosomal regions (Collantes-Fernández *et al.*, 2006).

It worth noting that such variations will contribute to genotype-phenotype relationships, more significantly, in the virulence and successful transmission strategy of *N. caninum* as has previously been confirmed in other apicomplexan parasites species including *Plasmodium falciparum*, *T. gondii*, *H. hammondi*, *Trypanosoma brucei* and *T. cruzi* (Gardner *et al.*, 2002; Milet *et al.*, 2010; Panunzi and Agüero, 2014; Walzer *et al.*, 2014; Cuypers *et al.*, 2017; Sharif *et al.*, 2017). It should be noted that SNPs markers gained significant attention over other genetic markers due to their many advantages (Vignal *et al.*, 2002; Raza, Shoaib and Mubeen, 2016). Firstly, this method can provide valuable data of much greater genetic diversity than other markers, which is significantly associated with specific genes and phenotypic changes. Furthermore, the effectiveness of SNP markers is higher compared to other genotyping methods, in particular after using the modern sequencing technologies of DNA and RNA among isolates (Vignal *et al.*, 2002; Dou *et al.*, 2012).

Previously, genetic diversity studies of *N. caninum* strains has been is limited compared to *T. gondii* isolates, which has been studied using several approaches among different geographical isolates (Schock *et al.*, 2001; Regidor-Cerrillo *et al.*, 2006, 2013; Donahoe *et al.*, 2015; Salehi, Gottstein and Haddadzadeh, 2015; Medina-Esparza *et al.*, 2016). By comparing data from several next-generation sequencing platforms, gene duplications were identified across the recent annotated assembly of the genomes (Adomako-Ankomah *et al.*, 2014; DeBarry and Kissinger, 2014; Lorenzi *et al.*, 2016). As noted in previous genomic studies, species specific CNVs that result in duplications or deletions in specific segments of the genome have been connection to gene family clusters (Barry *et al.*, 2003; Kissinger and DeBarry, 2011; Forrester and Hall, 2014; Leckenby and Hall, 2015). Furthermore, duplication or deletion were essential sources of genetic diversity in *N. caninum* and *T. gondii* strains, and within members of the same clonal lineage. A large number of the putative copies of variant and tandem duplicated genes were uniquely enriched in secreted pathogenesis determinants such as the rhoptry, surface antigen, micronemes and *Toxoplasma gondii* family proteins families which have been previously been linked to host cell attachment, parasite virulence, host range and phenotypes. It is expected that by comparing the genomic data from a wide range of strains the genetic features underpinning the phenotypes may be detected. Knowledge of what drives the phenotypic differences that have been noticed between distinct strains at the genomic level is crucial in aiding our understanding of determinants of genetic diversity in DNA sequences among species and the significant consequences of potential impacts on the *N. caninum* parasite from a range of perspectives including host range, pathogenesis, host-parasites interaction and phenotypes.

Large-scale studies frequently focus on the proportion of data that map to established reference genomes. With deep NGS sequencing methods, it is possible to identify a high quality dataset of unmapped genomic sequences that might contain significant genetic information that can be used to understand the genetic variation between the published reference genomes and the resequencing individual isolates. Despite advances in filtering software, there can still be a large fraction of unmapped whole genome sequencing reads that either falsely map to the reference or do not map (Van Der Weide *et al.*, 2016). Those unmappable reads can represent genomic regions that represent coding and non-coding regions of the genomes.

The main benefit of *de novo* reconstruction of the genomic data that were extracted from the unmapped reads is the ability to discover novel gene non-coding RNAs and new exons to gain evidence that *N. caninum* has variable genomes per strain to better understand it's pathogenesis, host range host-parasite mechanism and genetic diversity (Whitacre *et al.*, 2015). Indeed, the logical conclusion can be drawn from reconstruction is that it will to improve current gene annotation amongst varied reference and non-reference strains, most significantly by using NGS sequencing technologies with deep enough depth to assess the function of those putative strain - specific sequences.

## 4.2 Aims of the chapter

To our knowledge at the start of this investigation of SNP records, none were available in the public databases from multiple resequencing of strains of *N. caninum*. These findings of polymorphic genes, based on pairwise comparison, will allow us to compare directly varied individual isolates using the SNP marker to decipher the effect of the mutations more precisely as well as the frequency of SNPs clustered in specific regions within and between genomes in distinct strains of *N. caninum*. In addition, we investigated the genes that were enriched in copy number of variations to study the role of CNVs in the genome of *N. caninum*. Furthermore, it may be revealing evidence for genomic plasticity and evolution in the genes that coded for proteins. This work had multiple aims:

1. To identify sequence divergence level in the multiple strains, particularly in the genes that have high rates of SNPs by generally looking at genetic diversity level amongst the three distinct isolates collected from three geographical locations.

2. To study the different types of SNPs and predict what may be the functional consequences in the different genomes. This was done by looking at the specific and conserved regions in sequencing data of three populations with distinct biological features to pinpoint the significant genes causing the variations to investigate intra-species diversity.

3. To analyse the reads not mapping to the *N. caninum* reference genome for novel gene content. These predicted genes may offer additional insights into the biology of distinct strains of *N. caninum* and may contribute to our understanding of the epidemiology of neosporosis.

## 4.3 Results

### 4.3.1 Data generation and sequence read alignment of the *N. caninum* isolates to the entire *N. caninum* Liverpool reference genome

Raw sequence reads for *N. caninum* strains *NC-Bahia*, *NC1* and *NC-Liverpool* were generated according to the optimised workflow that was used by the CGR as described in Chapter 2. Contaminations were removed from the reads, identified from Metagenomics Phylogenetic Analysis software (MetaPhlAn) see section 2.12.1. After passing the filtering process, high quality clean reads were successfully extracted that contained only strain specific sequences per sample. Based on an estimated genome size of 59,103,012 base pair for the reference genome *N. caninum* strain *Liverpool* the coverage was ~8x. An alignment process was done of data from the three *N. caninum* strains to the recently published *N. caninum* Liverpool complete reference genome as explained in Chapter 2 in order to reduce the errors in the current reference genome that might have an influence on the next downstream analysis. Sequencing quality for all three isolated parasites was considered carefully in the subsequent steps of detection SNP and identification of the genes that were under selection in different strains of *N. caninum.*

The sequencing process yielded 176,999,394 reads, 75,343,042 reads and 61,923,976 reads of raw data in *NC-Bahia*, *NC-1* and *NC-Liverpool* respectively. The statistics for sequencing data quality in the three isolates is shown below in Table 4.1. The overall alignment rate per isolate can be observed in Table 4.2 from the data generated after sequencing. Large discrepancies were noticed in the proportion of read categories as we expected. The absolute number of total reads, mapped reads and unmapped reads between the three sequenced strains differed. As shown in Table 4.2, the mapping yielded 22,193,721 reads, 51,895,839 reads 53,921,425 reads of mapped reads to the updated reference genome in *NC-Bahia, NC-1* and *NC- Liverpool* strains respectively. A much greater number of successful mapped reads was identified in the *NC-1* strain (96.9%) reflecting the low number of mismatches allowed in the mapping tool used in this study.

As seen in Table 4.2, there was a decrease in mapping percentage for the *N.C-Liverpool* (73.77%) compared with the effective mapping rates in *NC-Bahia* and *NC-1* strains. The mean coverage was actually higher in the *NC-1* strain than the other two isolates reflected by the average coverage per chromosomes at 122.83 x compared to the other two strains.

The most striking result to emerge from the data was seen in chromosome II with the highest mean coverage of the three isolates (see Figure 4.1). Comparing the *N. caninum* strains, the lowest mean coverage was observed in V, XI and VIII chromosomes *in NC-Bahia, NC-1* and *NC- Liverpool* strains respectively. The whole genome sequencing data (see Figure 4.1) obtained from our study indicated a significant difference between the total proportions of unmapped reads from the *NC-Liverpool* and *NC-Bahia* data when mapped to the reference. Interestingly, the highest number of unmapped reads was observed in the *NC- Liverpool* resequencing isolate totalling 14,144,925 million (26.23%) then *NC-Bahia* with 12.3%. All the comparisons were made against the reference genome previously submitted by Welcome Trust Sanger Institute for *N. caninum Liverpool;* which assembled into 585 scaffolds with an N50 of 359 kb and an estimated genome size of 59.10 (Mbp).

**Table 4.1:** Summary of sequence read data generated by paired-end sequencing with 150 bp average length of reads of three *N.caninum* isolates; paired-end Sequencing ( R1: Read 1 forward), R2; Read 2 reverse) and (R0: single-reads with one end).

| Isolate | Raw reads (reads) | Trimmed reads (reads) | R1 forward | R2 reverse | R0 reads |
|---|---|---|---|---|---|
| ***N.C-Bahia*** | 176,999,394 | 174,918,562 (99.8%) | 86,420,656 | 86,420,656 | 2,077,250 (1.18%) |
| ***NC-1*** | 75,343,042 | 74,014,391 (98.2%) | 36,786,410 | 36,786,410 | 4,41,571 (5.96%) |
| ***NC-Liverpool*** | 61,923,976 | 61,259,284 (98.9%) | 30,408,397 | 30,408,397 | 4,42,490 (0.972%) |

**Table 4.2:** Statistics summarising the read data mapped for the three isolates; *NC- Bahia, NC-1* and *N.C-Liverpool* after aligning the reads to the NC-Liverpool reference genome.

| **Data type** | ***NC- Bahia*** | ***NC-1*** | ***NC- Liverpool*** |
|---|---|---|---|
| **Total reads** | 22,193,721 | 51,895,839 | 53,921,425 |
| **Total mapped reads** | 19,469,216 | 50,287,739 | 39,776,500 |
| **Percentage of total reads mapped to reference (%)** | 87.72 | 96.9 | 73.77 |
| **Unmapped reads** | 2,724,505 | 1608 | 14,144,925 |
| **Percentage of total reads unmapped to reference (%)** | 12.28 | 3.1 | 26.23 |
| **Mean Coverage (x)** | 45.67 | 122.83 | 96.88 |
| **Mean Mapping Quality** | 56.71 | 56.61 | 56.89 |
| **GC Percentage (%)** | 52.42 | 54.61 | 54.35 |

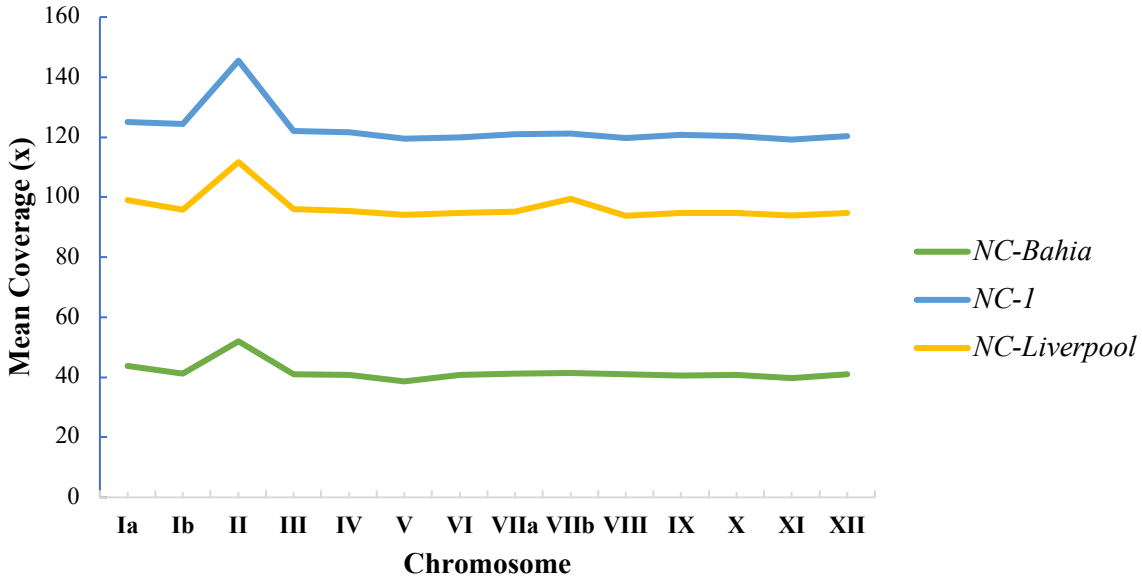



**Figure 4.1:** The summary statistics of mean coverage per chromosome of *NC-Bahia*, *NC-1* and *NC-Liverpool* strains generated by Qualimap tool v.2.2.1. The maximum mean coverage was in chromosome II in all samples. Blue line indicated the *NC-1* strain, green line *NC-Bahia* and the yellow line indicated the *NC-Liverpool* strain (see section 4.3.1).

### 4.3.2 SNP analysis

#### 4.3.2.1 Comparison of SNPs rate in three strains of *N. caninum*

SNPs were called and filtered by GATK pipeline to compare the genetic variation and determine the divergence levels within the three *N. caninum* strains. The level of uniqueness was determined using VCF-Compare and VCF-stat. Figure 4.2 shows the uniqueness and shared SNPs between each strain. This analysis revealed that there was a significant difference in the number of SNPs per strain and that the distribution of the unique SNPs was significantly different in the pattern of clustering per chromosome among distinct groups. Over the entire genome, a total of 10,729-14,021 and 6,697 SNPs were identified in the DNA sequences of *NC-Bahia, NC-1* and *NC-Liverpool* respectively (Table 4.4). We observed that the *NC-1* strain had a high coverage of 122.83x with a variant rate of 1 SNP every 5,497 bp (see Figure 4.3). Within the *N. C-Liverpool* resequencing sample, the low number of the SNPs and high percentage of similarity was expected due to the high similarity between the resequencing sample and the reference genome sequence that is publicly available (Table 4.3). As can be seen from our data, the variant rates were significantly lower in *NC-Liverpool* with 1 SNP every 8777 bp (99.4%) (Figure 4.3).

Inter-strain comparison revealed that there is a high degree of similarity, with 4038 SNPs conserved between the three strains, which accounts for 36.7%, 28.8% and 60.3% of the total SNPs that were called in *NC-Bahia, NC-1* and *NC- Liverpool* respectively. As a consequence of this high conservation, a low percentage strain specific SNPs was expected, which was 21.1% and 30.6% in both *NC-Bahia* and *NC-1* compared to the uniqueness percentage noticed in *NC-Liverpool* strain of only 13.1%. From our three - way comparison shown in Figure 4.4, SNPs that overlapped between *NC-Bahia* and *NC-1* total 4164 conserved SNPs. This contrasts with 264 shared between *NC-Bahia* and *NC-Liverpool* and 1519 shared between *NC-Liverpool* and *NC-1*. The strain with the highest level of uniqueness was the *NC-1* strain with 4295 strain-specific SNPs. For each strain, significant differences were robustly demonstrated in distribution of SNPs in contigs and chromosomes per strain. The majority of contigs have a varied number of SNPs per contig consisting of 15.4%, 11.5% and over 20 % of total SNPs called in *NC-Bahia, NC-1* and *NC-Liverpool*, respectively.

**Table 4.3:** The number of SNPs estimated in all *N. caninum* strains were called by using GATK3 that identified difference from the reference genome of *N. caninum Liverpool*. Here, it has been showed that there was a diversity per strain. The highest count of SNPs was found in *NC-1* strain and the lowest count was identified in *NC-Liverpool*.

| **Data type** | ***NC-Bahia*** | ***NC-1*** | ***NC-Liverpool*** |
|---|---|---|---|
| **Total number of SNPs** | 10,729 | 14,021 | 6,697 |
| **% SNPs in coding DNA sequence (CDS)** | 7.3% | 8% | 5% |
| **Number of nonsynonymous SNPs** | 999 (65%) | 1,434 (65%) | 378 (68%) |
| **Number of synonymous SNPs** | 513 (35%) | 774 (35%) | 171 (33%) |

**Table 4.4:** Summary of SNP calling generated by GATK3 tool. In this table, the data above shows the very high degree of similarity between the three strains. SNPs were generated after mapping to the reference genome *NC. Liverpool* available from (www.ToxoDB.org).This was done by VCF- Compare after filtering the data using final filtered draft of VCF file per strain.

| Sample | Site unique in isolates | %Unique in isolates | Site shared with other isolates | %Sites shared with other isolates |
|---|---|---|---|---|
| ***NC-Bahia*** | 2260 | 21.1% | 4038 | 37.6% |
| ***NC-1*** | 4295 | 30.6% | 4038 | 28.8% |
| ***NC-Liverpool*** | 873 | 13% | 4038 | 60.3% |

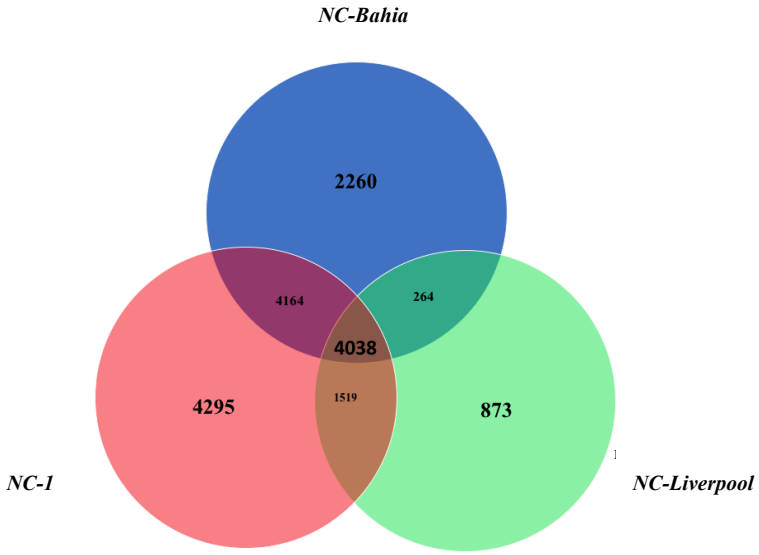



**Figure 4.2:** Venn diagram of unique SNPs comparison between the three strains. Blue circle indicates *NC-Bahia* strain, *NC-1* in red and in green was *NC-Liverpool* strain. The shared SNPs between the three samples are also illustrated. Each circle gives the total number of SNPs for each condition.

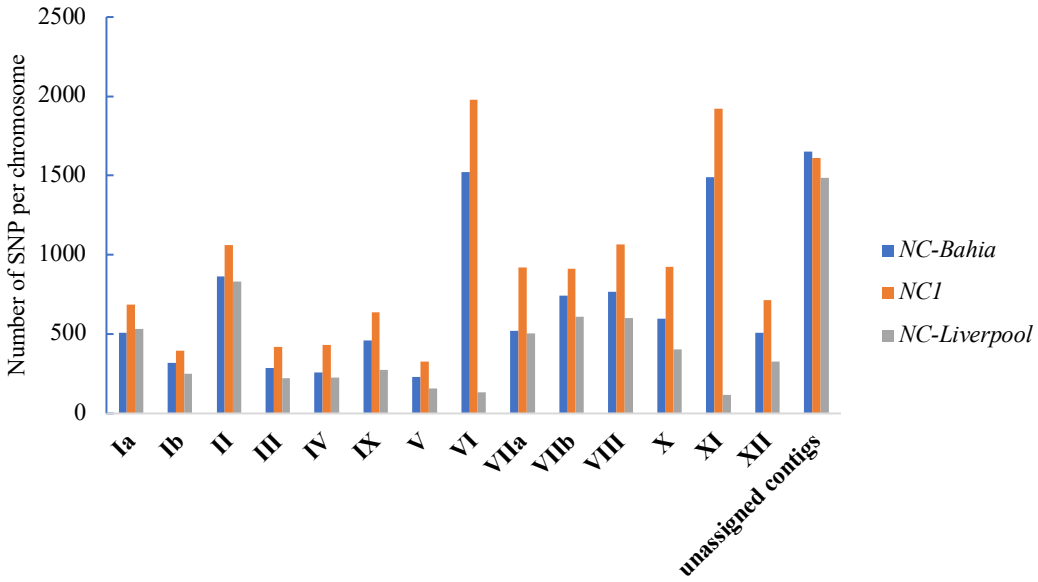



**Figure 4.3:** SNP number by chromosome in the three strains. The actual SNP counts are shown in three colours, blue: *NC-Bahia*, orange: *NC-1* and grey: *NC-Liverpool.*

#### 4.3.2.2 Representation of the SNPs within the data

The unique SNPs were plotted based on the genotypic change they cause. Transversions, A> C, C< A, G >T,T > G and A >T and transitions C >T, T > C, G >A and A > G generally occur unevenly with transition being a more common type of point mutation than a transversion event, which was expected in terms of occurrence. Transversions are less likely to produce a difference in the amino acid sequence than transversions and would thus remain as a silent SNP (Figure 4.6). It has been noticed that there was an unequal distribution of SNPs by looking in depth at to the density of SNPs in different chromosomal locations per strain. It is important to know which region of the chromosome was highly diverse, and more significantly, whether there is a correlation between those unique locations and the abundance of the SNPs that cluster in specific regions.

We next set out to look at the distribution of SNPs in each strain. As can be seen in Figure 4.9,4.11 and 4.13. The highest concentrations of SNPs were found in the subtelmeric regions. By looking at the density of SNPs unique to *NC-Bahia* and *NC-1*, it has been shown that two chromosomes (VI and XI) had a high number of SNPs across their length with 1523 and 1489 SNPs/1000kb in *NC-Bahia* and 1977, 1929 in *NC-1*. In chromosomes VI and IX for both isolates, the majority of SNPs were clustered at the beginning of the chromosome with 40% and 14% out of the total variants obtained for each. Comparing the SNPs pattern for chromosomes, VI and IX , we found that there was a further cluster in chromosomes VIII near to the chromosomal ends. The SNPs rate per chromosome was estimated as one SNP per 8777 bp that were called over full length genome of 59Mb. Clusters of patterns SNPs significantly varied based on the high density of SNPs across the chromosomes and also per specific region of the chromosome. Interestingly, chromosome XI showed a high content of SNPs in the centre region of the chromosome as we noticed in *NC-Bahia* and *NC-1* strains suggesting that the XI chromosome might have an impact on the strain-specific differences. By zooming into those regions using the Artemis Visualizing genomic tool, it was shown that there was accumulation of repetitive regions with high number of SNPs across the three genomes.

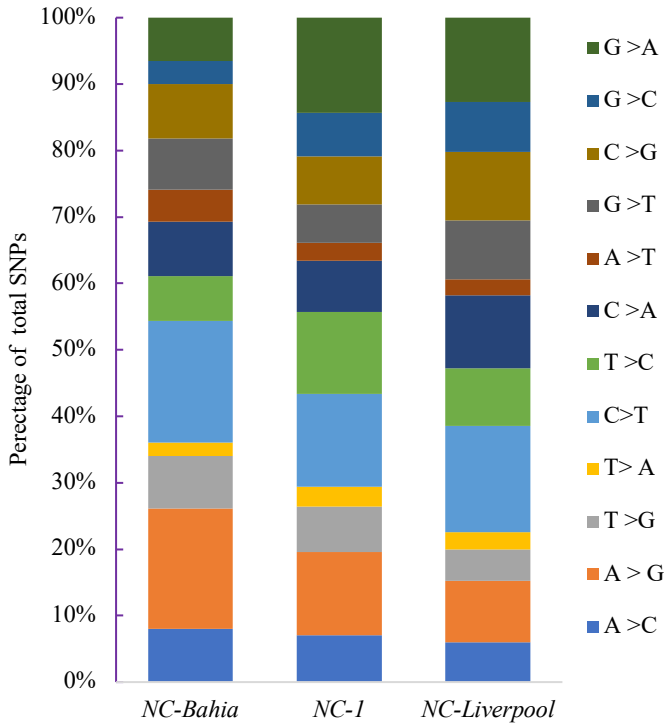



**Figure 4.4:** The percentages of SNPs and their corresponding mutation. This presents only small differences in the number of SNPs called between the three samples.

### 4.3.3 Investigating the frequency of genes that contain SNP within each strain

All the frequencies of SNPs per strain were calculated. Statistically significant differences were found in the frequency of the SNPs per gene obtained from our data sets. We additionally assessed the SNPs in the coding versus the non-coding regions of the genomes. The genes with the 10 highest unique SNP frequencies were included. In total, 1,113 genes contained less than 5 SNPs per gene as well as three isolates. Some of the polymorphic genes contained more than 50 SNPs (see Appendix A; Table A.1). Based on the current SNP analysis, there were additional SNPs found in unassigned contigs (UACs) with unknown function in all three isolates, as shown in Table 4.4. According to our results, *NC-1* strain has the highest number of SNPs. What stands out in this study was the dominance of chromosomes VI, VIIa that have the highest proportion of SNPs, suggesting that the genes are involved in the phenotypic changes might be located on these specific chromosomes. Within this list of genes, there was an Endonuclease/exonuclease/ phosphatase family protein, with a high number of SNPs that was commonly present in all isolates. This protein is important in signalling and phosphoric ester hydrolase activity. Additionally, a putative nucleoporin FG repeat-containing protein with 19 SNPs might be consider a divergent factor due to the unique structure of repeats.

The same trend of frequency was noticed in *NC-Bahia* strain although slightly different from *NC-1*, more precisely, in the number of SNPs per gene annotated and in the percentage of the genes that contained SNPs. More than half of the genes were mutated (63%) out of the total genes that contains SNPs. The majority of genes were grouped as unknown functions (hypothetical and conserved). There was a high degree of similarity in terms of gene annotations between *NC-Bahia* and *NC-1*. Interestingly, the highest abundance of SNPs were found in SRS genes that clustered in some specific chromosomes and regions. These included strain - specific genes that might alter effects with conservation in some hypothetical and conserved noticed in *NC-1* and *NC-Bahia*.

### 4.3.4 Investigation of genomic diversity within different sequence classes

We then went on to look at the predicted functional consequences of the SNPs in coding regions. Based on the impact prediction, the SNPs were categorized into four impact groups, low, moderate, modifier and high impacts. The impact predictions were varied depending on the function of protein products. The SNPs with low impact that unlikely change the functionality of the protein produced such as synonymous SNPs. The SNPs with modifier impact were affected the noncoding genes in both UTRs, intergenic SNPs and in putative regulatory sequences. In addition to those impacts' prediction, the SNPs with moderate impact was noticed in the nonsynonymous SNPs which causing a harmless effect on protein functions. However, the SNPs with high impacts that have a highly effect on the protein functions such as stop gained and frameshifts by causing change in the amino acids by distributing stop codons that cutting length of protein then causing large alterations and lose of protein functions. As we can see from Figure 4.5, most SNPs annotated across all three *N. caninum* strains caused a modifying impact, accounting for greater than 90% of the SNPs in each strain. The next largest category was the modifier impact SNPs accounting for no more than 6% of all SNPs annotated across three strains. This was followed by low impact SNPs with no greater than 3%.

Our results showed (see Figure 4.6) that the greatest number of SNPs were predicted to be located in intergenic region followed by downstream and upstream parts of the annotated genes across all the three strains. Only 7% ,8% and 5% of SNPs were located in exons and 9%,10% and 6 % were found in intron sequences in *NC-Bahia, NC-1* and *NC-Liverpool* respectively. As we described earlier, there was a significant decrease in the percentage of the SNPs that were distributed in splice site donors, acceptors and regions. Based on the SNP types, the SNPs differed in the three genomes. From Figure 4.7, it has been noticed that the individual SNP types were that predicted to have the highest percentage of effects were in intergenic regions located between genes. These accounted for approximately 36%, 33% and 50% out of the total SNPs per type in *NC-Bahia, NC-1* and *NC-Liverpool*, respectively.

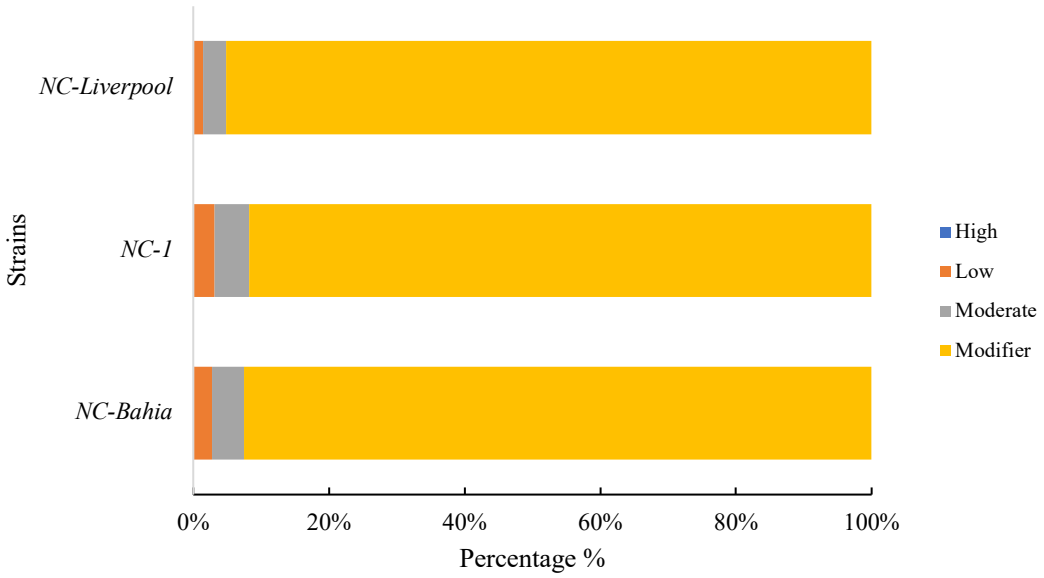



**Figure 4.5:** The number of SNPs per impact in three isolates of *N. caninum*. This categorises the SNPs into severity groups, low, moderate, modifying and high.

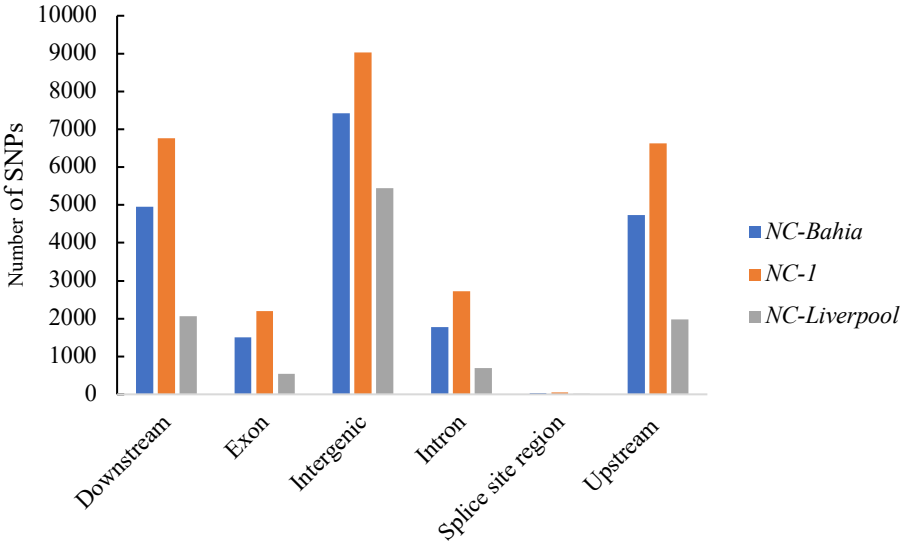



**Figure 4.6**: The number of SNPs per region in three strains of *N. caninum*.

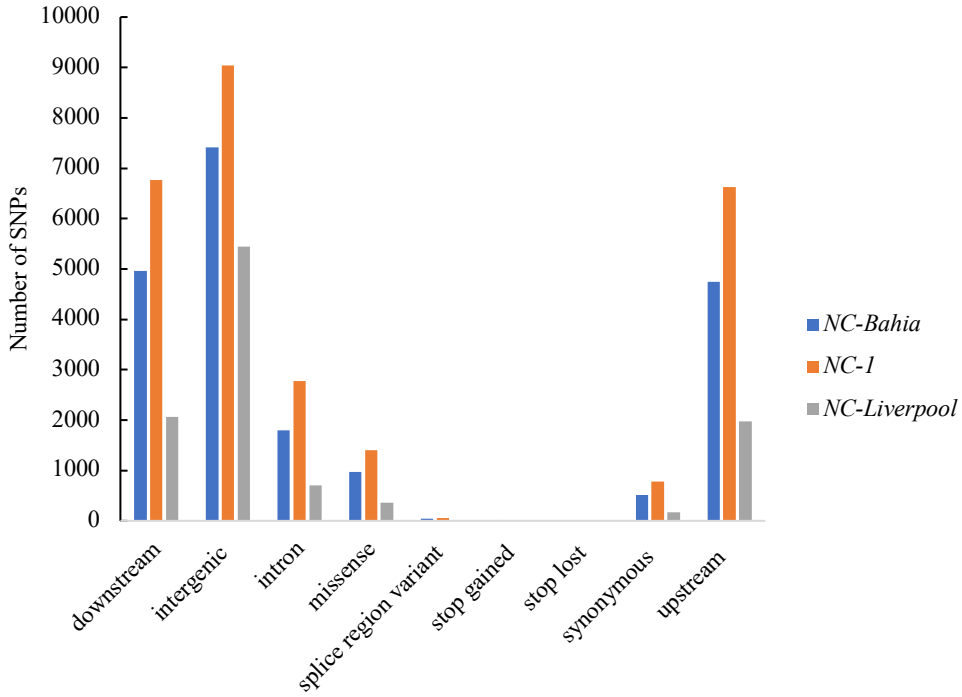



**Figure 4.7:** The number of SNPs per type in the three different strains of *N. caninum.*

### 4.3.5 High impact SNPs unique in three strains of *N. caninum*

In this study, we comprehensively catalogued the key genes that containing these high impact SNPs of *NC-Bahia*, *NC-1* and *NC-Liverpool* and classified them into distinct functions or subfamilies. In total 21, 39 and 15 SNPs with high impacts were predicted in *NC-Bahia, NC-1* and *NC-Liverpool* strains, respectively (see Table 4.5). Those SNPs residing in the genes are concisely explained below. The high impact terms were shown different degree of severity that generated some multiple functional effects per gene and not restricted to stop lost, stop gained and start lost functional effects. Based on the SNP high impacts analysis **(see Appendix; Dataset of chapter four (B),** there was one gene (NCLIV_02100), which encodes an SRS domain containing which is conserved in all three strains with one high impact SNP.

By looking at the unique SNPs per strain, it seems that there were some genes that have more than one high impact SNP including stop lost and stop gained changes which have highly disruptive effects due to premature stop codons in the sequences of the proteins products. In addition, our analysis demonstrated, that the *NC1* strain encodes more diverged genes than the other isolates. It has been shown that there was a putative protein kinase (NCLIV_0333570) belonging to the protein kinase-like family (PK-like), several of which have been previously confirmed in *N. caninum* and other apicomplexan members as coccidian virulence associated rhoptry kinase factors annotated as ROP46. Additionally, we identified one member (NCLIV_065460) where the presence of a SNP predicted a start lost effect belonging to the protein kinase known as the CMGC group, that contains four protein kinase families involved in different cell cycle signalling and regulatory process (Talevich and Kannan, 2013a).

To study patterns of diversification and functionalization of other genes and gene families with the high impact SNPs, we systematically identified further expansion of another superfamily that contains two significantly enriched genes belonging to the SAG -related surface antigen family that is located in chromosome VIIa with one high impact SNP per gene.

Analysis of the high impact SNPs in *NC-Bahia* strain revealed one conserved hypothetical protein had two high impact SNPs in chromosome XII with no orthologues to other species and of unknown predicted function. We also identified one gene annotated as AGAP005082-PA, (ID: NCLIV_056900), which is orthologous to the *T. gondii* gene TGME49_313630, which has three domains PF12624: Chorein_N, N-terminal region of Chorein, a TM vesicle-mediated sorter. This indicated that the full-length protein is a transmembrane protein with a presumed role in vesicle-mediated sorting and intracellular protein transport as well as homology to PF066050: Protein of unknown function (DUF1162) according to the Pfam database.

Not surprisingly, *NC-Liverpool* strain has the lowest number of predicted high impact SNPs. It has been found that the SNPs were clustered in chromosome VIIa in two different positions in the SRS genes (NCLIV_020092 & NCLIV_020100); these genes are also found in *NC-1* strain. Both genes are highly conserved and have one high impact SNP causing premature stop codon. The similar effects of those SNPs might be reflected in a high degree of relatedness between the *NC-1* and *NC-Liverpool* strains due to the conservation between them. Collectively, most of the functional effects in all three strains were stop gained leading to premature stop codons, otherwise known as nonsense SNPs and usually resulting in non-functional protein products; this was observed in *NC-Bahia* and *NC-1* strains. However, *NC-Liverpool* showed a high frequency of stop lost effects that are usually noticed in terminator codons. Additionally, the genomic locations of high impact SNPs were clustered in specific chromosomes such as VIIa, XI, and VIII. This might be correlated with variations between strains in terms of phenotype.

**Table 4.5**: The number of genes with predicted high impact SNPs in each of the three strains of *N. caninum.*

| Strain | Number of predicted high impact SNPs | Number of genes |
|---|---|---|
| ***NC-Bahia*** | 21 | 18 |
| ***NC-1*** | 39 | 35 |
| ***NC-Liverpool*** | 15 | 11 |

### 4.3.6 Investigating the most diverged genes within each strain

The SNP data analysis highlighted that the ratio of missense to silent did not vary greatly between the three isolates (1.8-2.2). The highest number of non- synonymous (missense) SNPs were found in NCLIV_002450 in all three isolates with more that 50 SNPs of unknown function as mention earlier in section 4.3.2.3. Generally, highly effective nonsense SNPs were identified in most of the genes with high impact SNPs. However, most of the other synonymous SNPs have low impact. Interestingly, the variants classified as missense generally had predicted moderate impact (see section 4.3.2.3). Regardless of the SNPs type, the total number of coding SNPs in *NC-Bahia, NC-1* and *NC-Liverpool* varied. The total number of genes that contained SNPs were 3,638- 4,728 and 1,341 in *NC-Bahia, NC-1* and *NC-Liverpool* respectively. The details regarding missense, nonsense and silent SNPs were plotted in Figure 4.8.

Our finding shows that the greatest number of non-synonymous SNPs were observed in *NC-1* totalling 1434 SNPs, includes 1415 (missense) and 19 nonsenses SNPs in 682 genes each with one or more SNPs. It has been revealed that the higher proportion of SNPs were also significantly related in the large number of genes identified. In order to examine the most diverged genes per strain, we carried out a comparison between the group of genes. From the functional annotation of the genes that contain the variants, we noticed that the majority of diverged genes were annotated as hypothetical, conserved and surface -antigen proteins, supporting the early evidence outlined in section 4.3.2.3. The lowest number of genes contained SNPs (non-synonymous) was in *NC-Liverpool* which account for 10 % compared to *NC1*. It was interesting to note that 29 genes (21%) of the total genes were annotated as surface -antigen gene family such as SRS33 as we expected and had noted previously.

To determine the influence of these SNPs in the three isolates, we examined the list of candidate genes after identifying the pattern of SNPs. We showed that there was a dramatic enrichment of surface antigen proteins totalling 25 members out of the total genes with nonsynonymous substitutions. This finding illustrates that there was a significant expansion and diversification of polymorphic potential genes that might reflect new biological characteristics due to evolution and mutagenesis in the genome of this Brazilian strain.

Several kinds of proteins were identified including; WD-40 domain, MORN and leucine rich repeats, Endoclase, ELMO and ALGC families and kinase proteins that all play varied cellular and molecular functions as we will discussed later in GO analysis (see section 4.3.7). In the *Liverpool* strain, the genes with markedly more SNP belonged to the SRS genes family; the remaining genes encoded unknown functions, some of which were conserved across all the isolates. Collectively, our results revealed that there was SNP diversity related to a large number of SRS family genes that were reported previously in *NC-Liverpool* but not in the *NC-1* and *NC-Bahia* strains and we postulate that those genes may well contribute to host range restriction of *N. caninum.* The SNP density per 1000/kb in the three isolates of *N. caninum* and example genes containing SNPs were plotted in Figures 4.9-14 per strain across the 14 chromosomes.

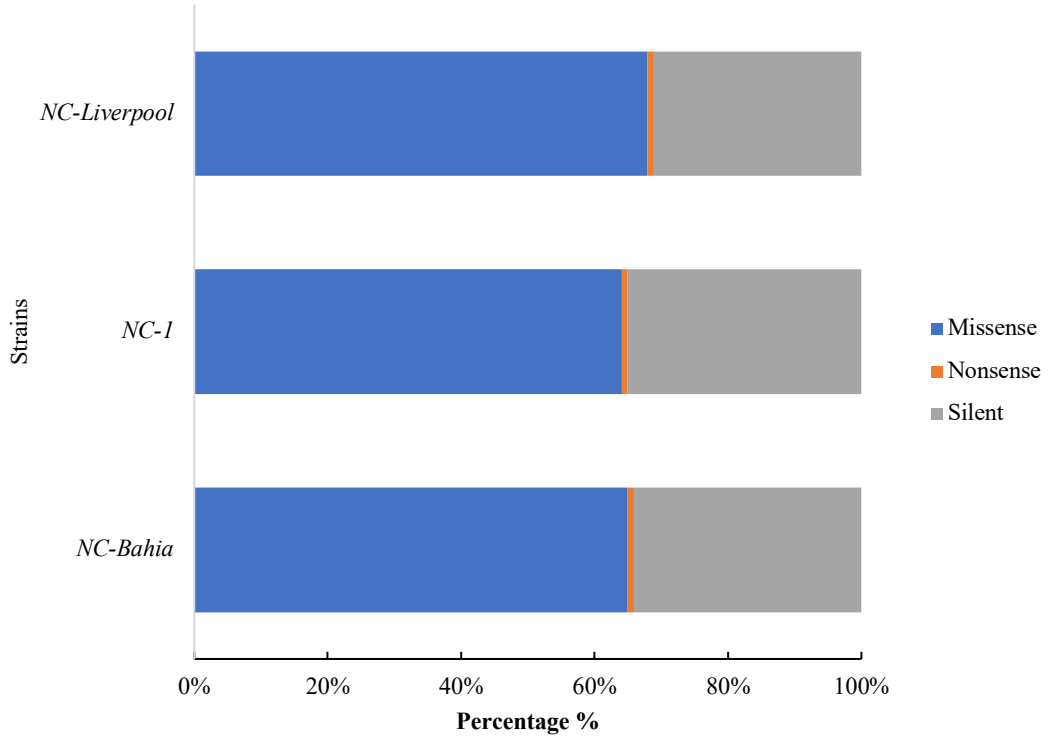



**Figure 4.8:** The number of SNPs by functional class in the three strains of *N. caninum.*

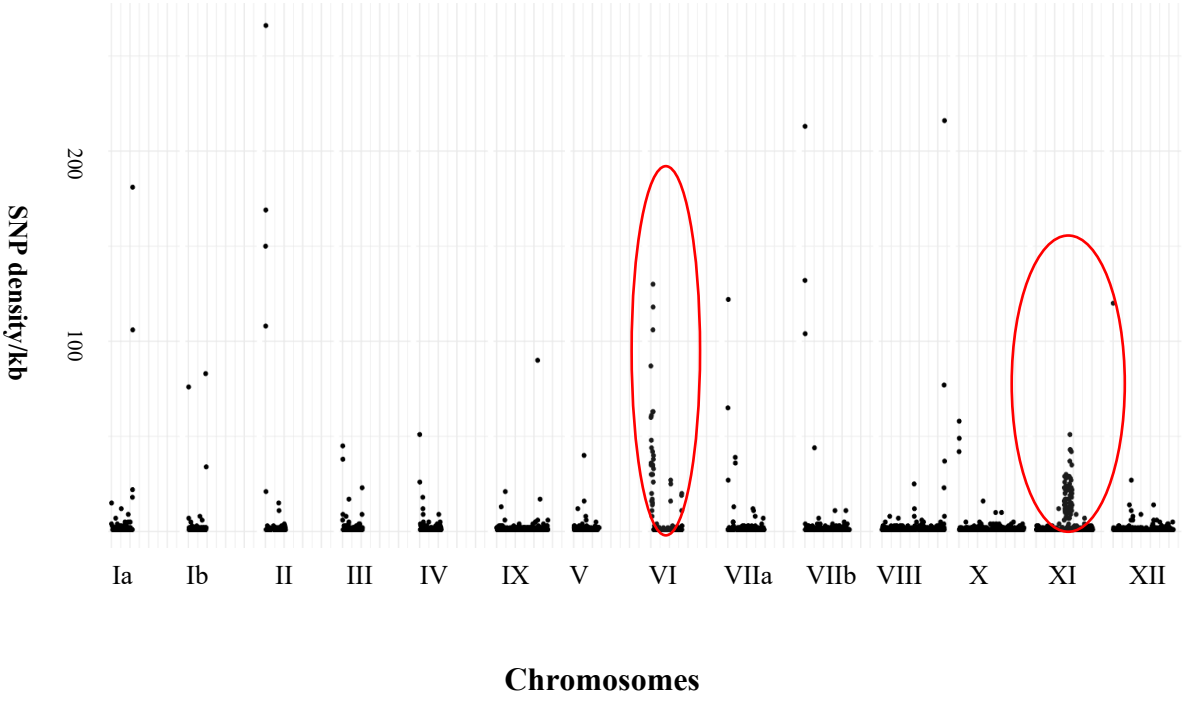



**Figure 4.9**: The SNPs density per 1000/kb in *NC-Bahia* strain; the red coloured circles showed the packed dense of SNPs (Clusters) in chromosomes VI and XI showing regions high density of SNPs. This Figure was generated using the R programme (gplots library https://www.rdocumentation.org/packages/gplots/versions/3.0.1.1)

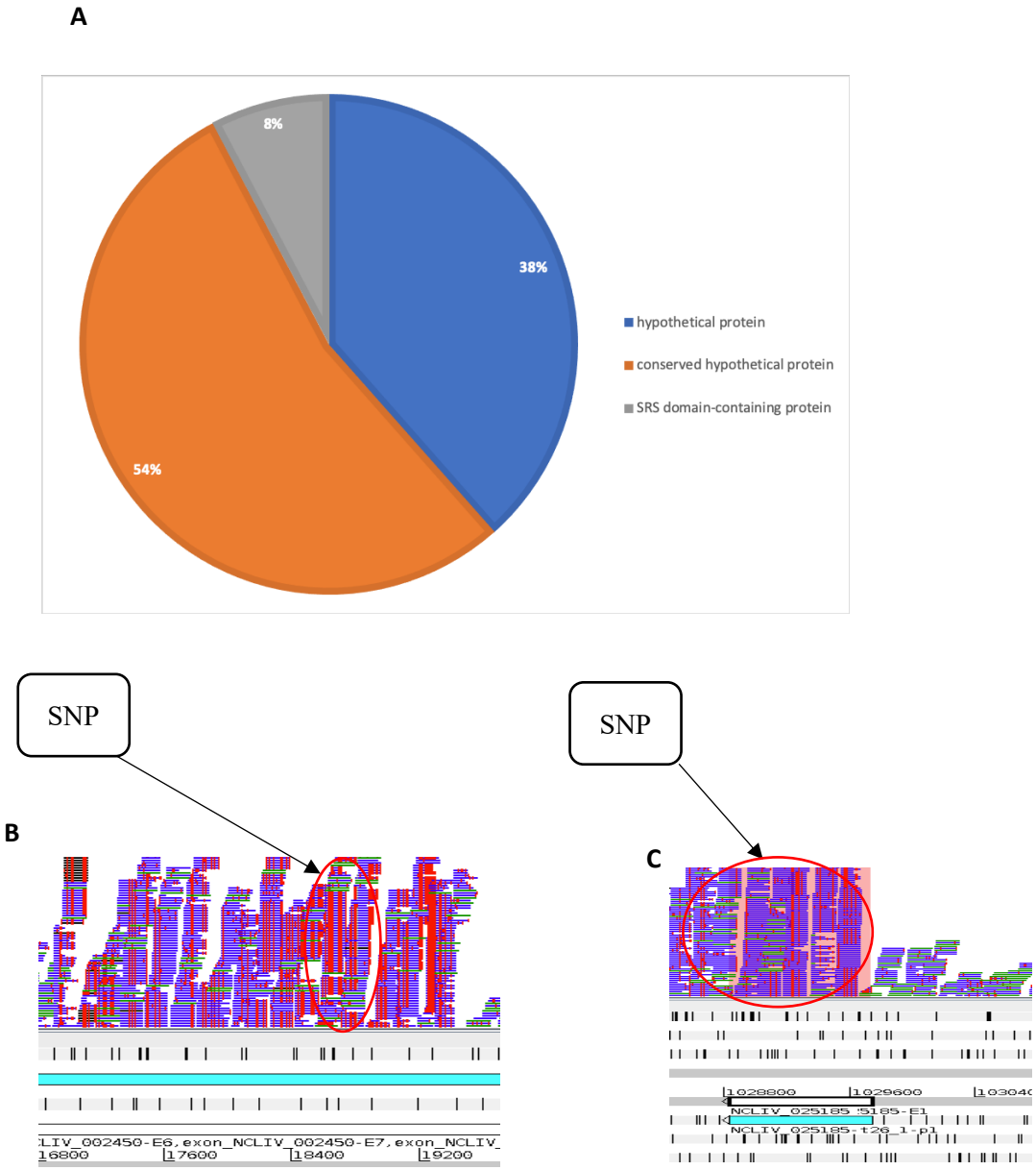



**Figure 4.10: A)** The frequency of non-synonymous SNPs within genes in *NC-Bahia*. **B)** Screenshot of the highest number of SNPs in a gene located in chromosome Ib (NCLIV_002450) that was found (conserved in all three samples). SNP graph shows red marks on the stacked reads that reflected the bases not matched with reference. The red vertical line indicates the real SNPs**. C)** Screenshot of the highest number of SNPs in SRS gene (NCLIV_025181) indicating the true SNPs. The screenshots in Figures B and C were visualized by using Artemis tool (https://www.sanger.ac.uk/science/tools/artemis)

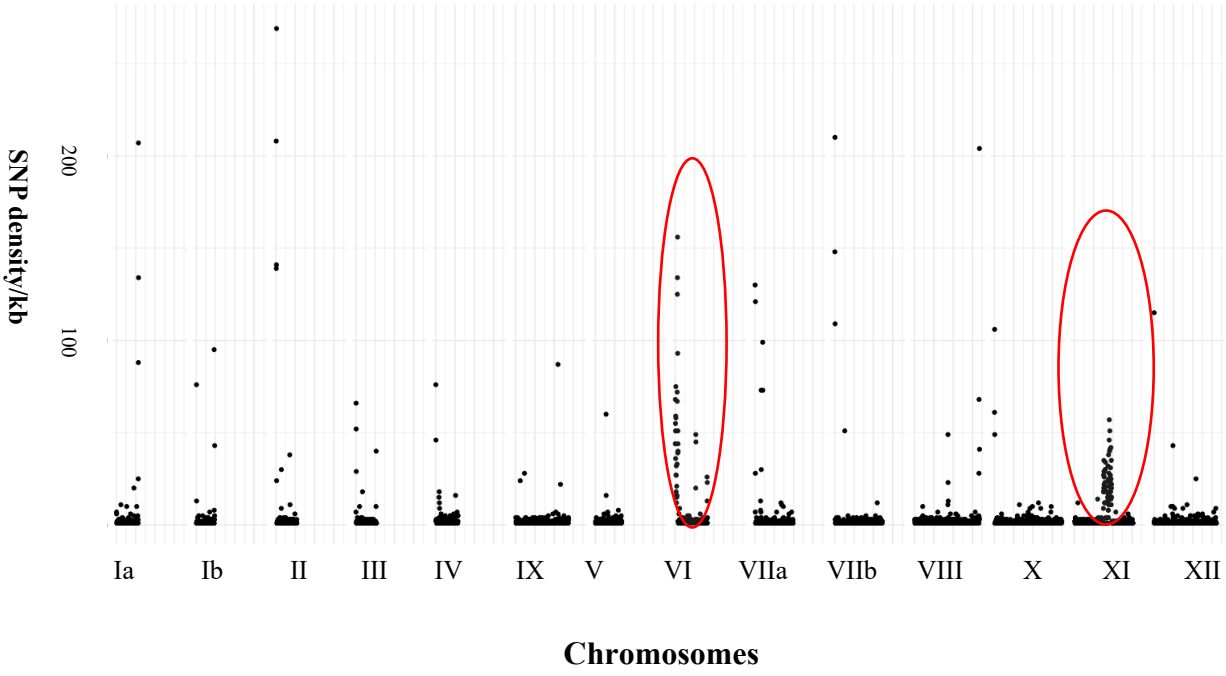
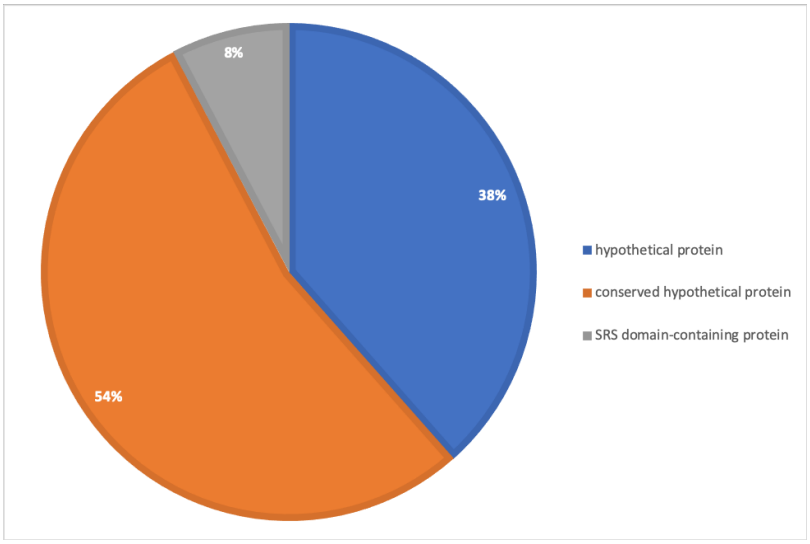



**Figure 4.11:** The SNPs density per 1000/kb in *NC-1* strain; the coloured circles showed the packed dense of SNPs (Clusters) in chromosomes VI and XI showing regions that have high density of SNPs. This Figure was generated using the R programme (gplots library https://www.rdocumentation.org/packages/gplots/versions/3.0.1.1)

A




B

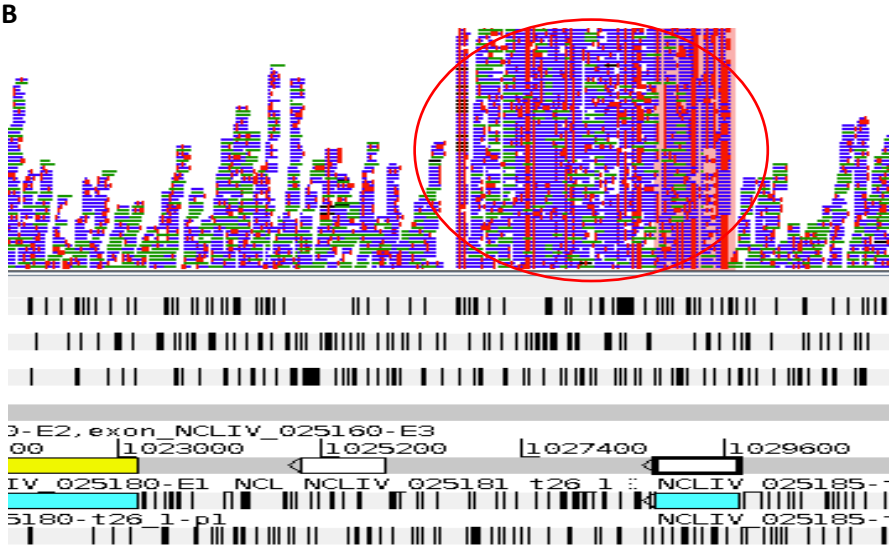



**Figure 4.12: A)** The frequency of nonsynonymous SNPs within genes in *NC-1* **B)** Screenshot of highest number of SNPs in SRS gene (NCLIV_025181). SNP graph shows red marks on the stacked reads that reflected the bases not matched with reference. The red vertical line indicated the real SNPs. The screenshot in Figures B was visualized by using Artemis tool (https://www.sanger.ac.uk/science/tools/artemis)

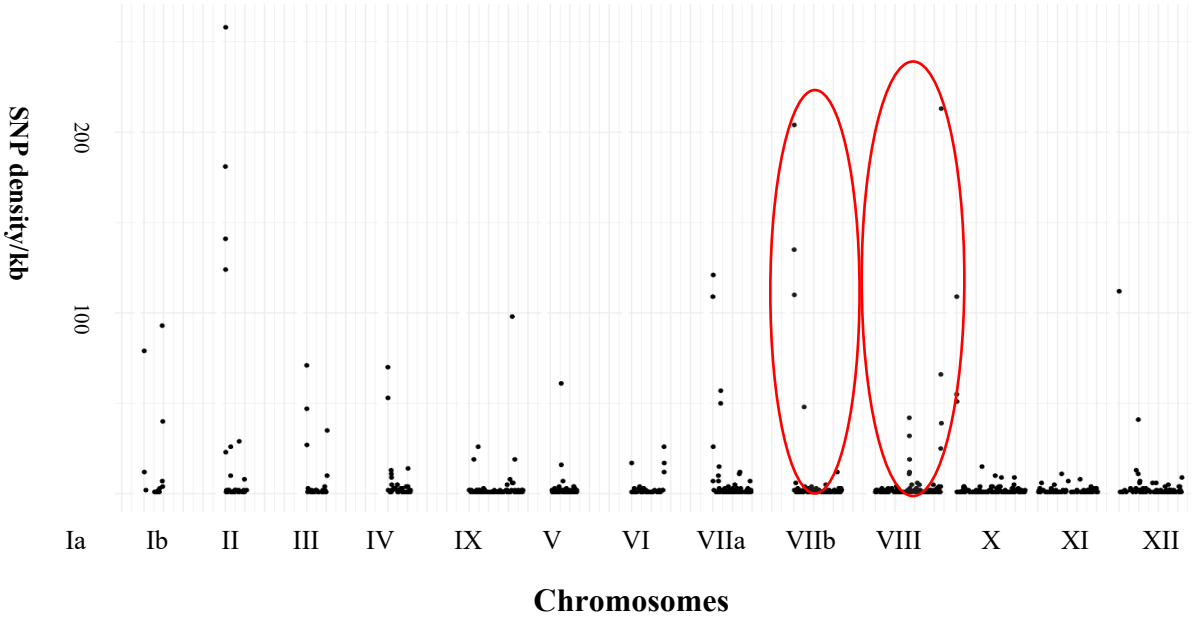
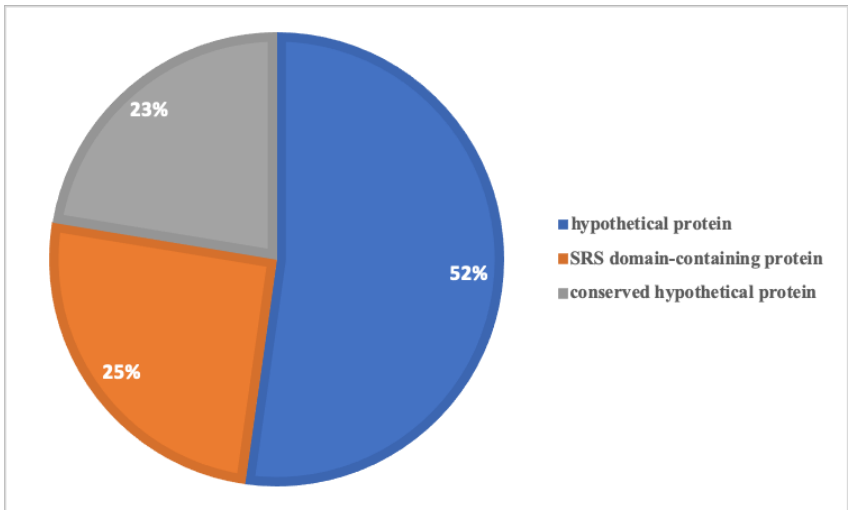



**Figure 4.13:** The SNPs density per 1000/kb in *NC-Liverpool* strain; the coloured circles showed the packed dense of SNPs (Clusters) in chromosomes VIIb and VIII showing the regions with high density of SNPs. This Figure was generated using the R programme (gplots library https://www.rdocumentation.org/packages/gplots/versions/3.0.1.1)

A




B

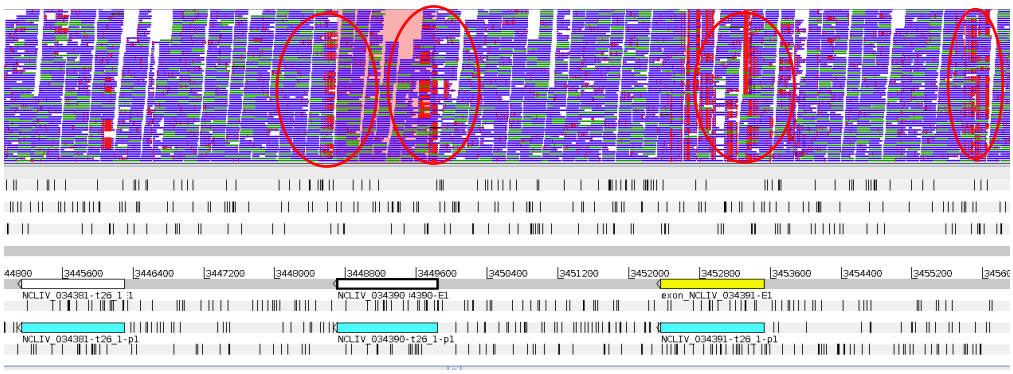


**Figure 4.14: A)** The frequency of nonsynonymous SNPs within genes *in NC-Liverpool* strain **B)** Screenshot of the highest number of SNPs in SRS genes located in chromosome VIII SNP graph shows red marks on the stacked reads that reflected the bases not matched with reference. The red vertical line indicates the real SNPs compared to the reference genome. The screenshot in Figures B was visualized by using Artemis tool (https://www.sanger.ac.uk/science/tools/artemis).

### 4.3.7 GO enrichment analysis

We next examined the lists of genes in the three strains of *N. caninum,* that contained SNPs in coding regions to determine whether GO term enrichment analysis could offer insight into explaining possible phenotype variations. For each strain, depending on the SNP impact categories (low, moderate, modifier and high) we used the GO term analysis to assign GO terms to the genes that were associated with the unique SNPs for each isolate. The ontology enrichment data were highlighted which terms have overrepresentation or underrepresentation compared to the rest by calculating the fold enrichment (the percentage of genes annotated to a GO term of interest divided by the percentage of the background). Enrichment was considered with P-value of less than 0.005 as explained in Chapter 2.

Characterisation of GO terms is important for understanding correlations between the specific pathways and phenotypic variations, particularity in the more virulent strains. The current Go terms annotations available via ToxoDB database for the reference organism *N. caninum* strain *Liverpool* include 3388 Go terms in three main ontologies. From the Figures 4.15 - 4.17, we can see that the *NC-1* strain enrichment in GO terms is significantly greater than for the other two strains. Each circle indicates a GO term, with the greater P value assigned. The greatest number of GO terms were assigned to modifier and moderate effect groups.

#### 4.3.7.1 Pathways enriched in genes containing predicted modifier impact SNPs

There were 16,10 and 20 GO terms assigned to the genes with modifying SNPs in strain *NC-Bahia, NC-1* and *NC-Liverpool* respectively. (see Figure 4.15). Notably, large number of Gene Ontology enrichment terms were noticed in the *Liverpool* strain. What stands out in Figure 4.15-C is the dominance of GO terms assigned to genes related to protein glycosylation modification processes. This was expected due to the presence of the surface antigen gene (SRS) family that was expanded in the *Liverpool* strain. It was also suggested that those genes found in the 20 clusters have the highest rate of SNPs with modified impact that strongly supporting the rapidly evolving of SRS leading to host invasion.

GO function assignments were enriched for proteins targeted to the membrane. As discussed earlier, it was noticeable that there was overrepresentation of a cluster concerning response to stimulus that might reflecting the ability of the parasite to protect their genomes under strong stress conditions in different host populations. It can be seen that there was a significant enrichment of GO ontologies terms in *NC-Bahia,* with 16 GO terms. From the GO enrichment analysis, most enriched GO terms with significant P-values were plotted in scatterplot in Figure 4.15 -A. The cluster representatives were given coloured bubbles to indicate the enrichments terms. In *NC-Bahia* strain, the most GO terms were enriched for general metabolic processes that affect several functions of the *N. caninum*. Also response to stress and response to stimuli that acting in the host -parasites functional such as secretion and production of enzymes that lead to changes in gene expression in virulent strains. Additionally, it might explain the mechanisms of transformation from tachyzoites to bradyzoites and *vice versa.*

It can be seen that there was a significant reduction in the number of GO terms in the *NC-1* strain involved in movement of cell and cellular components (GO:0006928) and carbohydrate catabolic process (GO:0016052). This pattern might reflect a combination of genes that were responsible for mechanical processes for specific secretion from micronemes (MICs) that penetratet to the host cell during host cell invasion, mechanism of movement by changing the structure during invasion. A clear link was observed here between highly enriched genes that contain these mutations and microtubule cytoskeleton as shown in Figure 4.15 B. Additionally, (GO:0006650) was also observed including glycerophospholipid metabolic process. This might suggest a direct effect in initial interaction with the host and regulation of many processes.

#### 4.3.7.2 Pathways enriched in genes containing predicted moderate impact SNPs

Most of the genes with SNPs that have a predicted moderate impact encode nonsynonymous SNPs that including missense that alters codons to produce another amino acid or nonsense substitutions that result in stop codons to produce truncated proteins with deleterious effects. To provide an overview of what particular pathways were enriched per strain. See Figure 4.16, all terms were assigned to 6,16 and 20 GO terms with fold enrichment more than 5 in *NC-Bahia*, *NC-1* and *NC-Liverpool* respectively. The greatest proportion of GO terms were observed in *NC-Liverpool*, followed by *NC-1* then the *NC-Bahia* strain with the lowest number of enrichment terms.

Looking at the frequency of GO terms in *Liverpool*, the majority of biological process are summarized from 20 enriched GO terms in Figure 4.16-C. More significantly, they include protein glycosylation and response to stimulus, implying roles in host-parasite interaction. It has been found that there were further significant biological process involved namely cell communication and signalling. The second highest proportion of enrichments were noticed in the *NC-1* strain with 16 GO terms that mapped to movement of cell and cellular components. This suggested that enriched shared genes may contribute to physical movement in the host - parasite interaction. Additionally, two GO terms were identified that are involved in regulation of protein catabolism and potassium transportation that support the hypothesis of the presence of some divergence between the virulent strain *NC-Liverpool* and the less or moderate virulent strain *NC-1,* as summarized in Figure 4.16-B. Fewer genes that caused moderate effects were observed in the *NC-Bahia* strain with only 6 GO terms as is evident in Figure 4.16 A. The clusters were significantly enriched in nitrogen metabolism and protein phosphorylation. Many of the genes in those groups belong to kinase genes that can have biological function in virulence.

#### 4.3.7.3 Pathways enriched in genes containing predicted high impact SNPs

The comparisons involving the three lists of genes that have high impact have been made to identify potential genes where their products have a severe functional effect. We presumed that those genes would be genes with associated pathways that might underlie phenotypic changes among strains. Figure 4.17 - A shows the GO terms in *NC-Bahia* that were involved in a variety of biological process related to response to stress and response to stimulus. One gene (NCLIV_049770) has shown a diverged domain known as PAN-domain that has a role in diverse biological processes by mediating protein-protein or protein-carbohydrate interactions. Based on the enrichment analyses, most of the Go terms in *NC-Liverpool* map to key biological metabolic pathways involved in RNA splicing. As significant reduction of GO terms was noticed in NC-1strain with primarily related to the DNA repair and damaged that have a significant function in nuclease activity and to keep the stability of the genome against any damage or changes in the DNA. Overall, our data showed there were differences impacting on all three strains in terms of SNP more importantly, pathways were correlated with the high frequency of high impact SNPs.

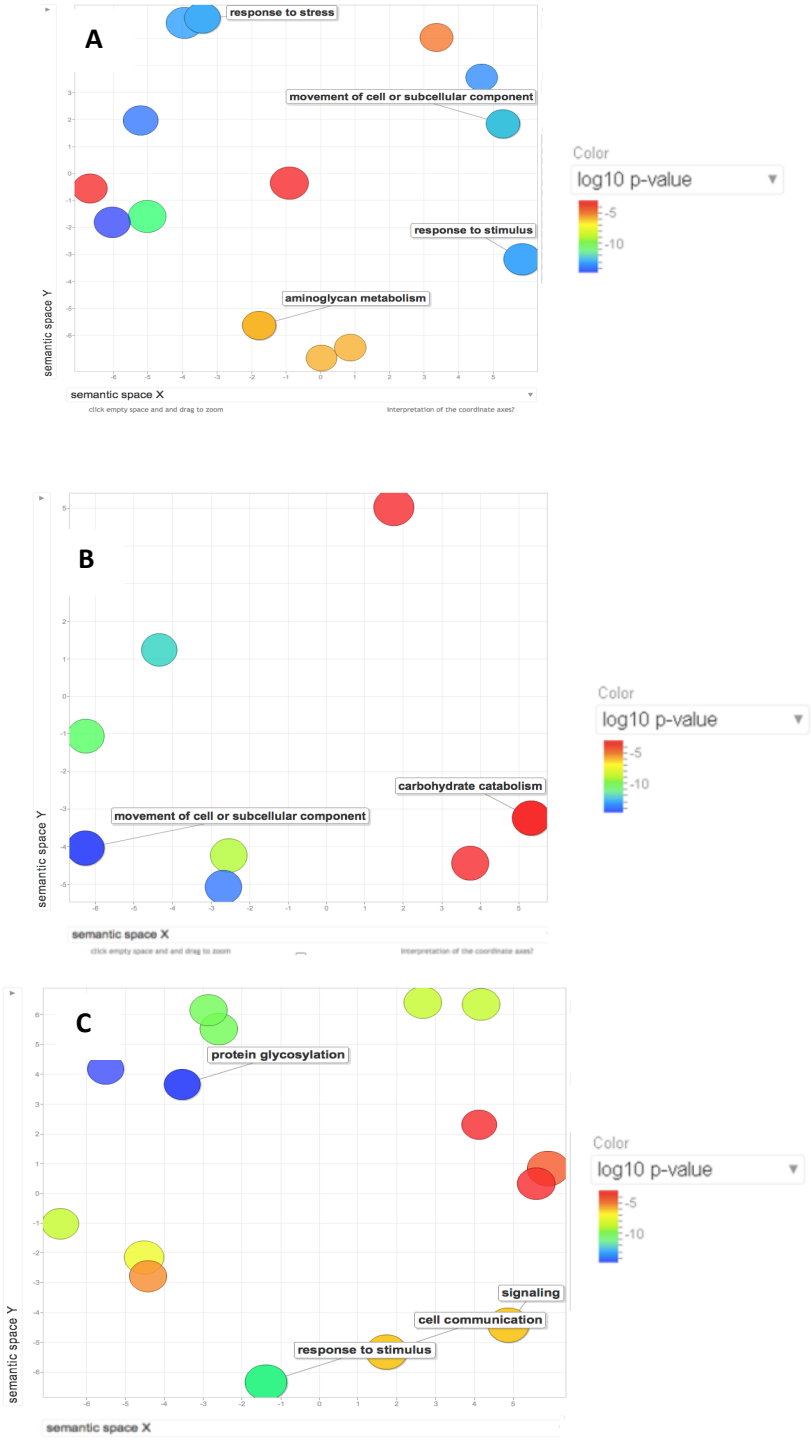



**Figure 4.15:** The scatterplot shows the cluster representatives (the GO terms remaining after the redundancy reduction) in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms' semantic similarities of SNP with modifying effect per strain **A)** *NC-Bahia.* **B)** *NC-1*. **C)** *NC-Liverpool* showing the cluster of genes representing biological process by REViGO software (http://revigo.irb.hr). The bubble colour indicates the p-value; size indicates the frequency of the GO term (bubbles of more general terms are larger).

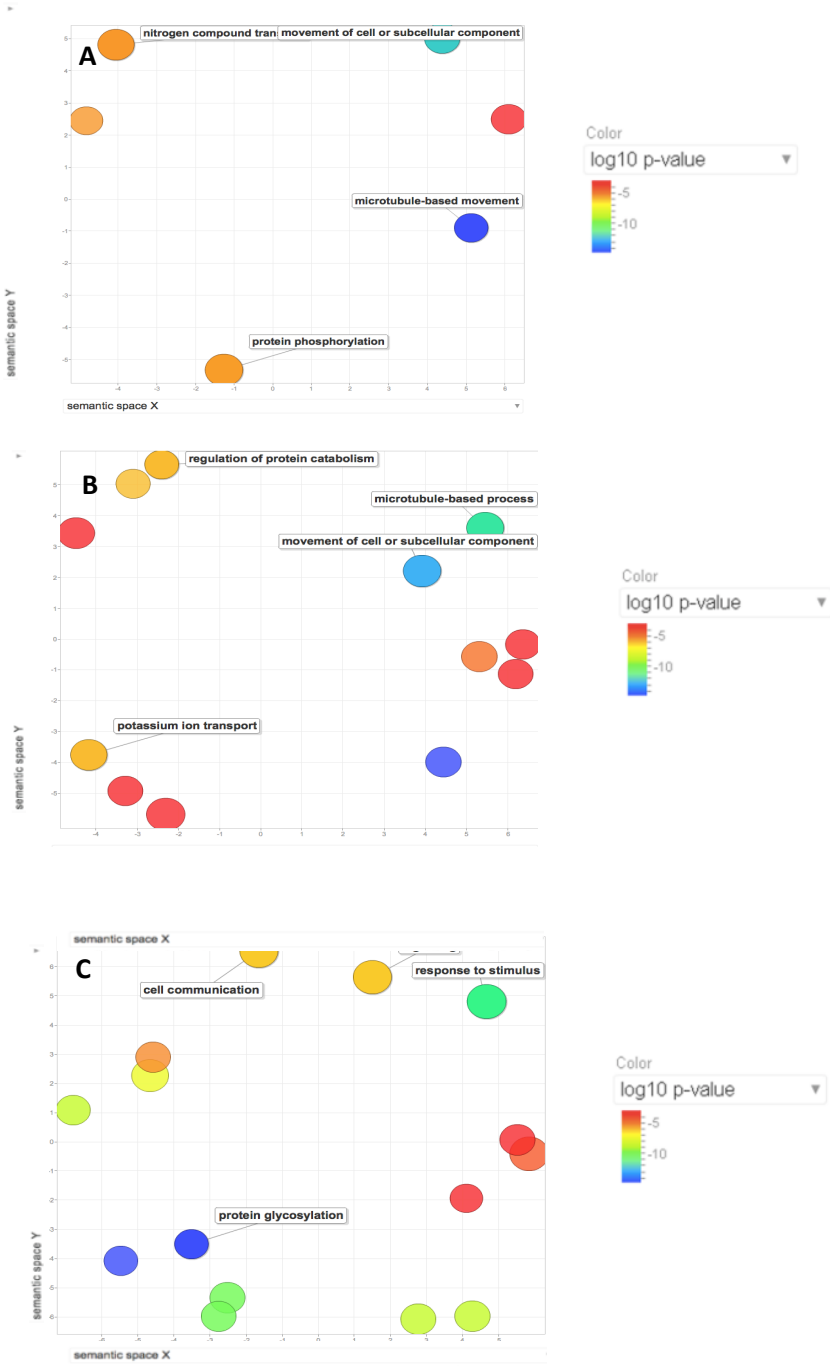



**Figure 4.16:** The scatterplot shows the cluster representatives (the GO terms remaining after the redundancy reduction) in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms' semantic similarities of SNP with moderate effect per strain **A)** *NC-Bahia.* **B)** *NC-1*. **C)** *NC-Liverpool* showing the cluster of genes representing biological process by REViGO software (http://revigo.irb.hr). The bubble colour indicates the p-value; size indicates the frequency of the GO term (bubbles of more general terms are larger).

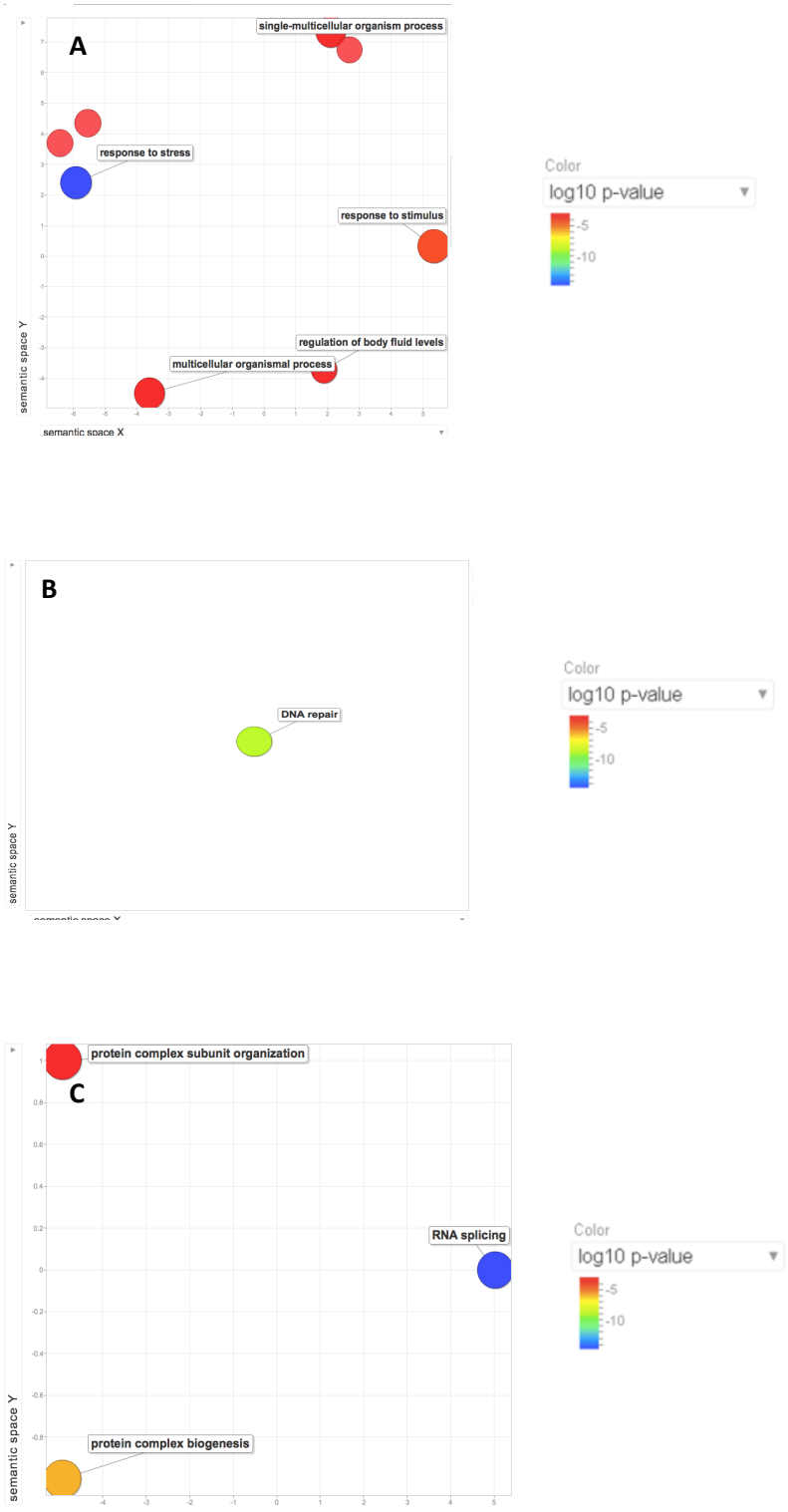



**Figure 4.17:** The scatterplot shows the cluster representatives (the GO terms remaining after the redundancy reduction) in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms' semantic similarities of SNP with high effect per strain **A)** *NC-Bahia.* **B)** *NC-1*. **C)** *NC-Liverpool* showing the cluster of genes representing biological process by REViGO software (http://revigo.irb.hr). The bubble colour indicates the p-value; size indicates the frequency of the GO term (bubbles of more general terms are larger).

### 4.3.8 Copy number variation (CNV) and the gene annotations of *N. caninum* strains

In addition to SNPs, variation in the copy number and the gene family content is an essential source of genomic diversity amongst *N. caninum* strains. As we mentioned earlier (see section 4.3.1), large differences in coverage of the three strains genomes of *N. caninum*, indicated copy number variation (CNV) between the strains. The frequency of CNVs, including deletions and duplications, varied among the three strain of *N. caninum*. Using the CNVentor tool package, the genomic regions that underlie the CNVs were identified and calculated per strains to determine whether particular gene/gene families were enriched and whether that variation might contribute to phenotypic changes and virulence within *N. caninum* strains. There was a significant association of the CNV with genomic locations and with the length of genes. We examined the outputs of the pipeline and identified and curated the full set of genes that underlie the CNV regions and compared the cluster positions and the degree of overlapping between the three strains in all 14 chromosomes.

In total we identified 40, 80 and 60 duplication in the *NC-Bahia*, *NC-1* and *NC-Liverpool*, strains respectively. Our analysis demonstrated that there was a dramatic increase in the number of duplications in the *NC-1* isolate, which had the highest coverage compared to the other two strains. We compared the number of duplications and deletions with their lengths in the three *N. caninum* isolates. However only a small number of duplications was noticed in the *NC-Bahia* strain. A large number of putative CNV genes was observed in all 14 chromosomes in *NC-1* (see Figure 4.18). Seventy percent of those genes are annotated as SRS domain-containing proteins while the remaining annotated as hypothetical proteins. Fewer genes were identified in *NC-Bahia,* totalling 20 putative duplicated genes. The vast majority of those genes belonged to the surface antigen-1-related (SRS) family. This reflected a significant enrichment of this specific gene family in the three *N. caninum* strains. Interestingly, it has been noticed that most of the duplicated genes are located in the telomeric regions of the chromosomes.

As we can see from Figure 4.19, the chromosomes X has the highest number of duplicated genes enriched with SRS proteins from the three isolates, the number and length of deletions was significantly higher than the duplications. A much longer length of deletion was identified in the *NC-Bahia* strain totalling 1,212 and 5900000 bp respectively. It has been reported that the highest deletion event occurred in chromosome XII located from 3967001-4031000 bp. Visual inspection revealed that the gene NCLIV_065330, which encodes a putative ATPase, was deleted. The lowest number of deletions was noticed in *NC-1* strain.

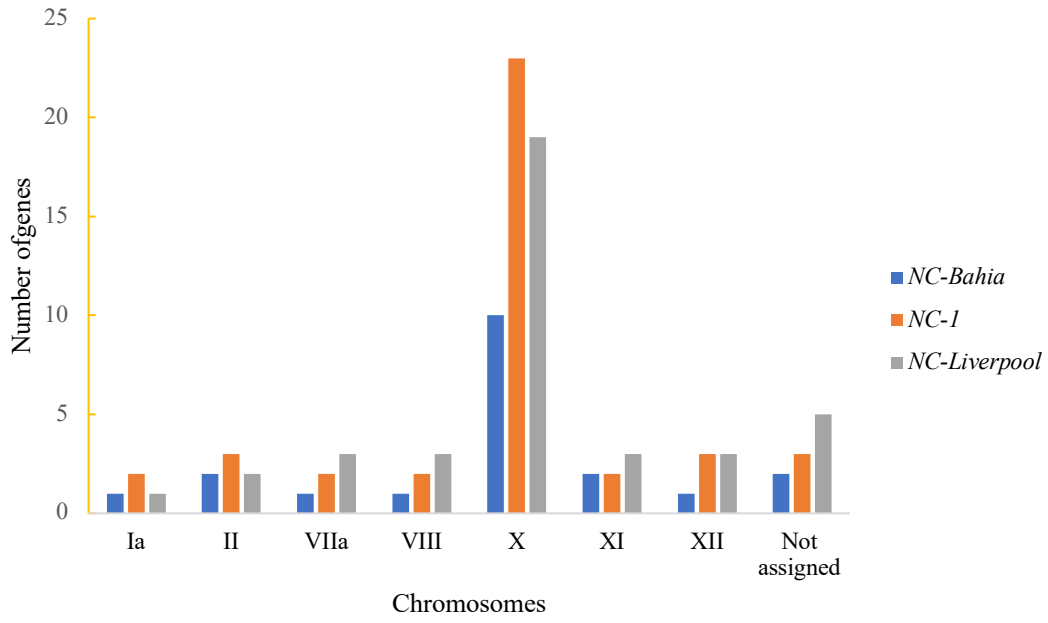



**Figure 4.18:** The distribution of putative duplicated genes that have more than one copy per chromosome in the three isolates of *N. caninum.* The chromosome X has so much duplicated genes enriched with SRS proteins (see section 4.3.8).

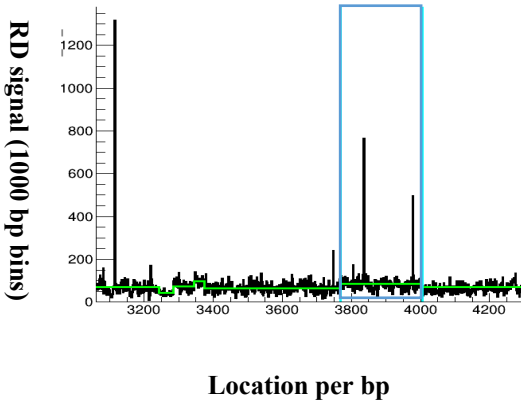



**Figure 4.19**: An example of large event of duplicated regions marked in blue rectangle in chromosome X of *NC-1* strain. The black histogram is indicated to the Read Depth (RD) signal for the fragment of chromosome X. Green line is partitioning by CNVnator programme (http://sv.gersteinlab.org/cnvnator/)

### 4.3.9 *De novo* assembly of unmapped reads analysis

When analysing the reads generated for each of the three isolates, we identified a subset of reads that probably originated from the mammalian Vero culture cell line and bacterial DNA and hence represented contamination of the original sample. To investigate this, the total number of reads per strain was entered into BlobTools to identify true sequences that were derived from the specific target genomes without any contaminants. Unmapped reads and contigs were subjected to a series of taxonomic annotation steps using the GC content of sequences and read coverage. The unmapped read sequences in the assembly were plotted and coloured based on the taxonomy assignment provided from command line parameters by the user. Based on those input sequences, there were many of sequences that originated from taxonomic groups other than Apicomplexan as can be seen in Figures 4.20 and 4.21. The majority of the sequences belonging to the Chordata phylum with total genome size of 4.34 Mb and 442 nt. Further taxonomic screening of this taxon revealed the GC proportion in this phylum was between 0.4 - 0.6 % forming a large cluster of orange points as shown in the Blopblot in Figure 4.20 and 4.21. This large contribution of Chordata genome, was derived from the host genome of the African monkey kidney cell line, known as Vero cells, based on BLASTn. In general, the coverage profiles varied across the different taxonomic groups including in the assembly and ranged from 0.1-1000X due to unequal depth in the genomes. The second source of contamination was the bacterial - derived reads which were also identified in the assembly due to experimental contamination. This was expected given the lower contents of GC %.

Out of the total data generated, 12.28% and 26.23% of the data remained unmapped after mapping to the updated reference genome for *NC-Bahia* and *NC-Liverpool* (see Table 4.2). After screening out the contaminating reads, the remaining reads were extracted and assembled *de novo* using SPAdes. In order to QC the assembly, we analysed these final assemblies via the Blobplot workflow again (data not shown). The visualisation of the assemblies per sample is shown below in Figure 4.22. In *NC-Bahia*, overall, the *de novo* assembly resulted in 425 contigs; 811513 bp total and with N50 of 2280 bp. The assembly output was assessed using the QUEST assembly assessment tool. There was a significant increase in the number of the contigs resulting in *NC- Liverpool* strain yielding a total of 560 contigs; 1287322 bp.

The N50 was 1823 bp. All assembly statistics were based on the size of contigs provided in Table 4.6 per strain The GC % content was relatively equal between the two strains, with 54.46 % and 54.08 % in *NC- Bahia* and *NC-Liverpool* respectively.

**Table 4.6**: The overview of assembly quality assessment for the genome assemblies per strain. Statistics were based on contigs of size >= 500 bp, unless otherwise noted (e.g: contigs (>= 0 bp) and total length (>= 0 bp) include all contigs).

| **Assembly statistics metrics** | ***NC-Bahia*** | ***NC-Liverpool*** |
|---|---|---|
| **Contigs** | 425 | 560 |
| **Contigs >=0bp** | 932 | 2254 |
| **Contigs (>= 1000 bp)** | 195 | 242 |
| **Contigs (>= 5000bp)** | 25 | 22 |
| **Contigs (>= 10000 bp)** | 4 | 7 |
| **Contigs (>= 25000 bp)** | 0 | 1 |
| **Contigs (>= 50000 bp)** | 0 | 0 |
| **Total length (>= 0 bp)** | 811513 | 1287322 |
| **Total length (>= 1000 bp)** | 525951 | 622988 |
| **Total length (>= 5000bp)** | 192125 | 243519 |
| **Total length (>= 10000 bp)** | 59292 | 135023 |
| **Total length (>= 25000 bp)** | 0 | 38537 |
| **Total length (>= 50000 bp)** | 0 | 0 |
| **Largest contig** | 20490 | 38537 |
| **Total length** | 691463 | 847818 |
| **GC (%)** | 54.46 | 54.08 |
| **N50** | 2280 | 1823 |
| **N75** | 1066 | 963 |
| **L50** | 72 | 90 |
| **L75** | 188 | 256 |
| **Mismatches** | | |
| **N's per 100 kbp** | 31.82 | 119.13 |
| **N's** | 220 | 1010 |

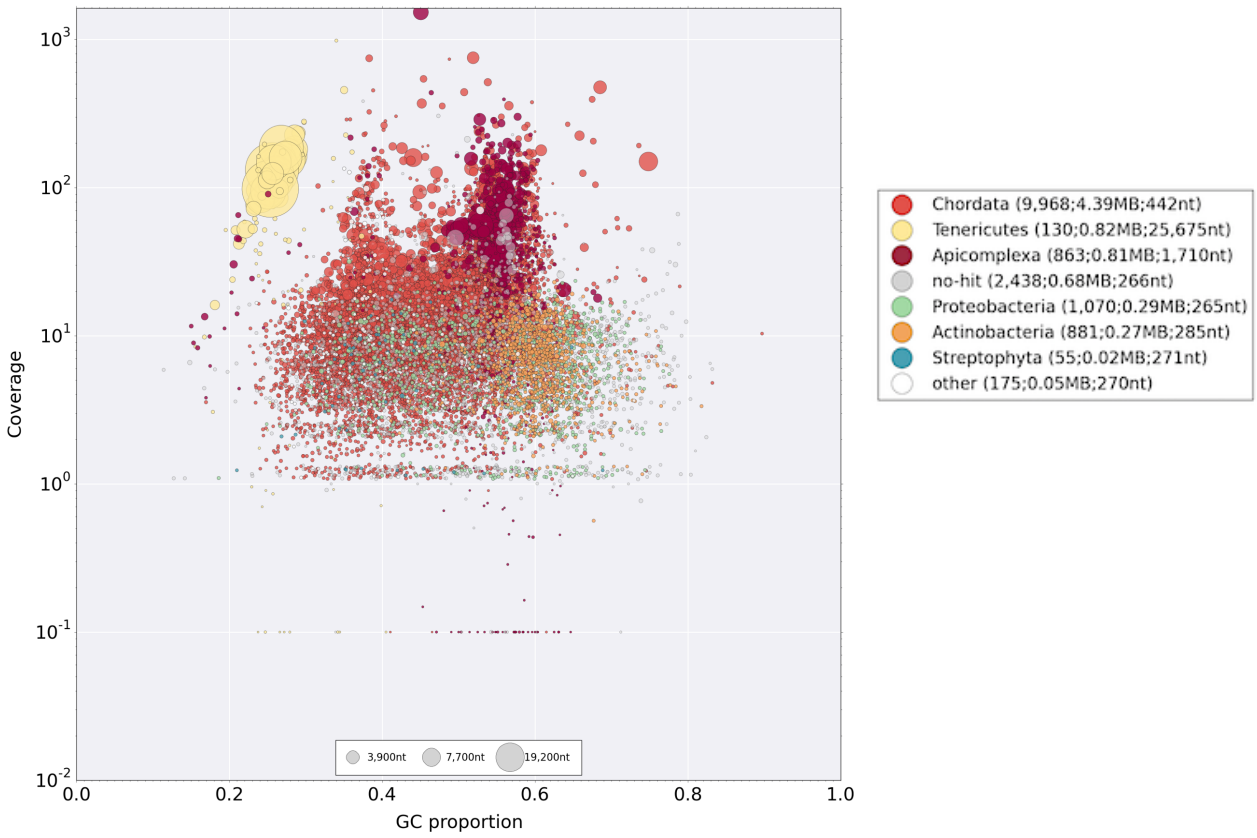



**Figure 4.20:** Blob Plot of the assembly depicted as coloured circles with ranks based on the taxonomic order, based on BLASTN similarity research provided from the tool in *NC-Bahia* sample (First round).

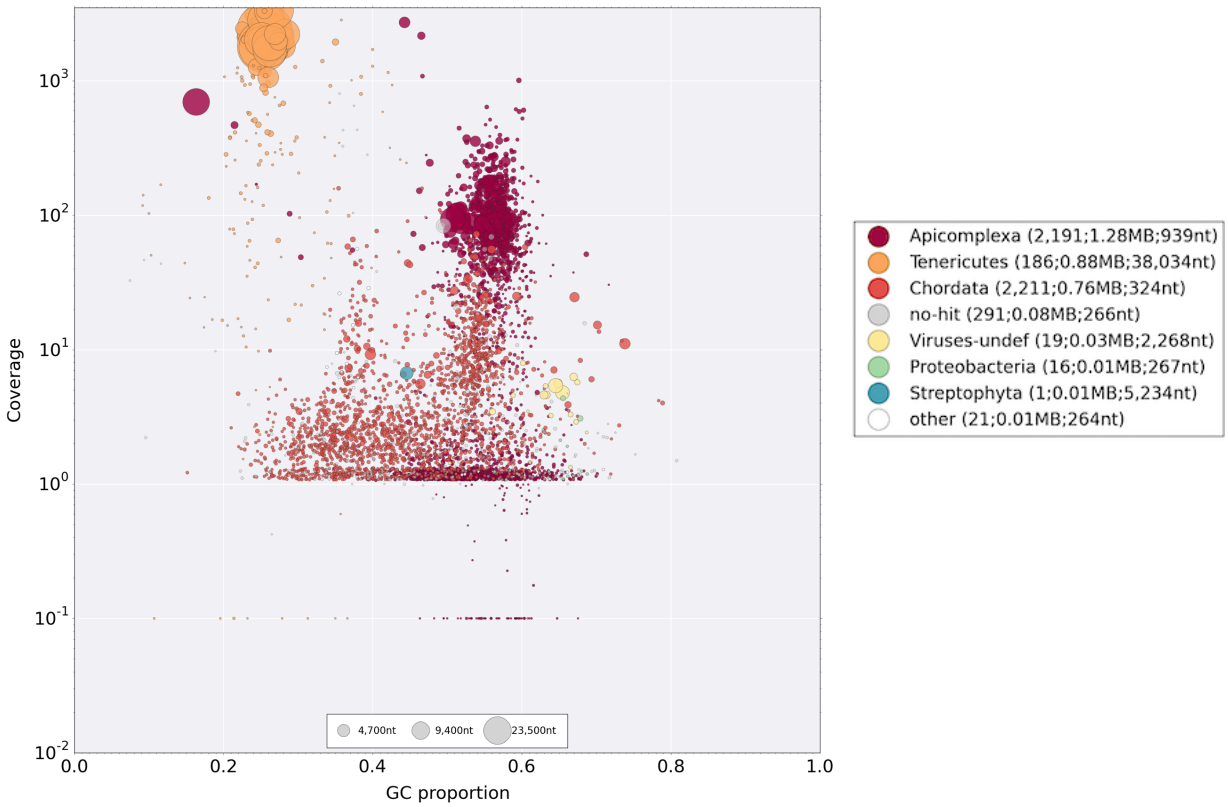



**Figure 4.21:** Blob Plot of the assembly depicted as coloured circles with ranks based on the taxonomic order, based on BLASTN similarity research provided from the tool in *NC-Liverpool* sample (First round).

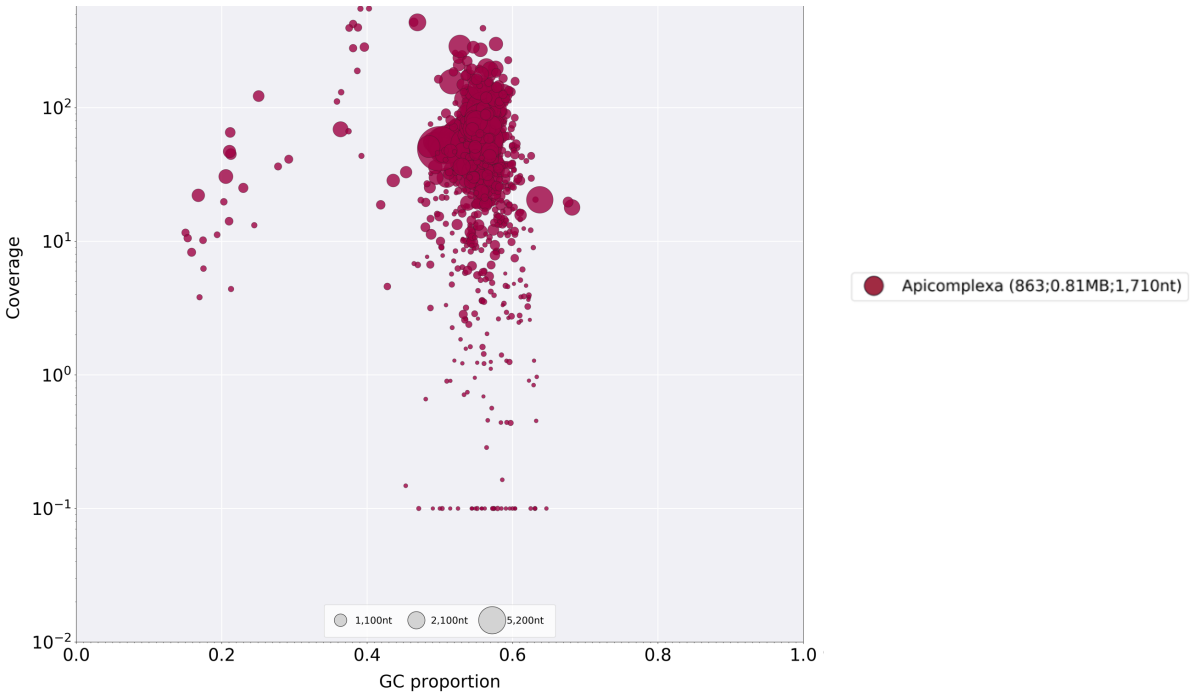



**Figure 4.22:** Blob Plot of the final purified assembly contigs that were extracted after removing contamination from the reads based on the taxonomic order identified using BLASTN similarity provided from the tool in *NC-Bahia* sample (second round).

### 4.3.10 Gene finding and annotation

We next went on to analyse the assembly of unmapped reads. Structural and functional annotation using the Companion tool identified 44 genes in the *NC- Bahia* and *NC-Liverpool* strains respectively that classified into coding and non-coding genes. A total of 23 genes were identified in resequencing of the *NC-Liverpool* strain and show significantly more non-coding genes with 18 genes, 17 tRNA genes and one rRNA, when compared to the *NC-Bahia* strain, as shown in the Table 4.7. The number of coding genes was lower than what we had in the *NC-Bahia* strain. From the Table 4.7, one can see that a considerably lower percentage of the overall GC% contents and more precisely, in the coding GC% in Liverpool strain. As shown in Venn diagrams in Figure 4.23, the shared and species-specific protein-coding of the target genomes of *NC-Bahia* and *NC-Liverpool* revealed that there was a significant number of singleton genes with no orthologues or paralogues in either species (see also below). The full list of species-specific genes is presented in Table 4.8. BLAST searches against the non-redundant protein database from NCBI using the protein sequences for all the 11 coding genes (6 in *NC-Bahia* and 5 in *NC-Liverpool*) identified no significant homology matches against the proteins in the NCBI database. Not surprisingly, there were numerous alignments to other close relatives of Apicomplexan such as *Toxoplasma gondii* and *Hammondia hammondi*. These results confirm novel gene content as compared to the previously released *NC- Liverpool* reference genome. In terms of the number of genes that have known functions, it has been found there were four genes, two in each sample, that can be ascribed a potential function based on OrthoMCL v1.4 cluster assignment and on BLASTP hits (see Table 4.8). In *NC-Bahia*, our cluster data clearly show that two newly identified genes; NEOSPORA_000006900 (GYF domain containing protein, putative) have an orthologue (ORTHOMCL232) with *N. caninum* and this cluster has 3 genes and 2 taxa; NCLIV_020970 NCLIV_027770 but with varied known functions. NEOSPORA_000006200 gene (Kelch motif/Galactose oxidase, central domain containing protein, putative) belong to the other cluster (ORTHOMCL270) that contains 3 genes, two with an 2 unknown function. In *NC-Liverpool*, there was a cluster (ORTHOMCL607) with has 2 genes and 2 taxa (NCLIV_039660) and the newly identified gene LIVERPOOL_000007100 was annotated as gf17338, related.

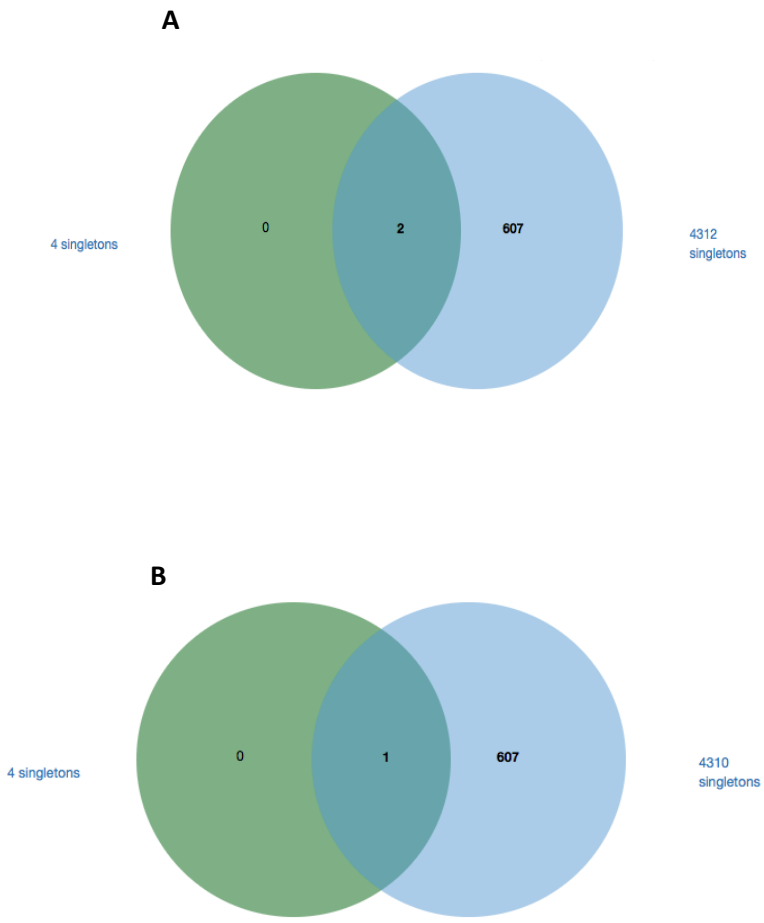



**Figure 4.23:** The Venn diagrams show shared and species-specific protein-coding gene clusters in the two target genomes; left, green **A)** *NC-Bahia* and **B)** *NC-Liverpool* versus the *N. caninum Liverpool* reference genome; right, blue. The number of singletons is shown in the outside of the diagrams.

**Table 4:7:** Genome statistics from *NC-Bahia* and *NC-Liverpool* strains based on the gene annotation results from Companion tool.

| Genome statistics | *NC-Bahia* | *NC-Liverpool* |
|---|---|---|
| **Number of annotated regions/sequences** | 1 | 1 |
| **Number of genes** | 21 | 23 |
| **Number of tRNA** | 14 | 17 |
| **Number of rRNA** | 1 | 1 |
| **Number of mRNA** | 6 | 5 |
| **Gene density (genes/mega base)** | 7.4 | 3.89 |
| **Number of coding genes** | 6 | 5 |
| **Number of pseudogenes** | 0 | 0 |
| **Number of genes with function** | 2 | 2 |
| **Number of pseudogenes with function** | 0 | 0 |
| **Number of non-coding genes** | 15 | 18 |
| **Number of genes with multiple CDSs** | 0 | 1 |
| **Overall GC%** | 54.43% | 54.03% |
| **Coding GC%** | 63.75% | 61.16% |

**Table 4.8:** The species-specific genes in the two strains of *N. caninum; NC-Bahia* and *NC-Liverpool.* The clusters of protein-coding genes in both strains were created by OrthoMCLv1.4 based on BLASTP hits.

| Strain | Gene ID | Gene Annotation | Comments |
|---|---|---|---|
| ***NC-Bahia*** | NEOSPORA_000005300 | hypothetical protein | - |
| | NEOSPORA_000005400 | hypothetical protein | - |
| | NEOSPORA_000006300 | hypothetical protein | - |
| | NEOSPORA_000007000 | hypothetical protein | - |
| | NEOSPORA_000006900 | GYF domain containing protein, putative | Orthologues to NCLIV_020970 conserved hypothetical protein NCLIV_027770 hypothetical protein |
| | NEOSPORA_000006200 | Kelch motif/Galactose oxidase, central domain containing protein, putative | Orthologues NCLIV_010741 hypothetical protein<br><br>NCLIV_014230 hypothetical protein |
| ***NC-Liverpool*** | LIVERPOOL_000006200 | hypothetical protein | - |
| | LIVERPOOL_000006300 | hypothetical protein | - |
| | LIVERPOOL_000007000 | SNARE domain containing protein, putative | - |
| | LIVERPOOL_000007200 | hypothetical protein | - |
| | LIVERPOOL_000007100 | gf17338, related | Orthologues NCLIV_039660 gf17338, related |

We next looked at whether potential protein domains could be ascribed to our newly identified genes. Our searches revealed two genes in *NC-Bahia* and *NC-Liverpool* had not been previously reported in *N. caninum* strains. Those novel genes belonged to two specific domains; GYF domain containing protein, putative and Kelch motif/Galactose oxidase, central domain containing protein. GYF domain (IPR003169) plays a critical importance in binding proteins. It has been proposed that the GYF domain could also be involved in proline binding. GO terms analysis of this gene (GO:0005515) revealed that the molecular function of this domain is protein binding. The length of the second gene was 338 bp which belonged to the Kelch motif/Galactose oxidase, central domain containing protein (PF01344). These domains were associated with various molecular functions as protein binding that were confirmed from several databases. Based on the protein sequence analysis, this domain is also widespread in other species but more importantly, in apicomplexan species such as *T. gondii* strains ME*49, P89 ,ARI,COUG* and *H. hamondi* with 89% of an identity that is known as leucine zipper-like transcriptional regulators.

As might be expected, the proportion of proteins varied due to diversity within genomes. In the *NC-Liverpool* strain, a putative the SNARE domain containing protein was identified in the resequencing sample. Interestingly, sequences containing the SNARE domain associated with secretory and endocytic trafficking functions due to their localisation in the intracellular membrane. From ToxoDB, five genes were identified with domain. Our novel gene LIVERPOOL_000007000 has this domain, so might play a significant role in vesicle-mediated transportation processes and protein binding. In the LIVERPOOL_000007100 gene annotated as gf17338, related has 139 bp in length and has orthologues with 63% identity to NCLIV_039660 that has been already annotated. It has been found that proteins that belong to this superfamily contribute to control of the stability of a group of inflammatory genes that might be important in pathogenesis.

## 4.4 Discussion

In this study we used whole genome sequencing to analyse the genetic heterogeneity in three isolates of *N. caninum*; these isolates originate from different hosts and different geographical regions. The comparative genomic analysis shows there were differences between the three strains by using SNPs and CNVs as tools for detection. To our knowledge, such a comparative genomic analysis based on the SNPs and CNVs between the three strains of this species has not been previously carried out. Varied amounts of sequence data were generated for *NC-Bahia*, *NC-1* and *NC- Liverpool*. Whilst this resulted in different levels of coverage across the genomes, the differential coverage was not associated with a significant increase in the number of calling SNPs. Although limited DNA was collected from the slowest growing isolate *(NC- Bahia*), we expected that might have an influence on DNA quantity and due to this, low coverage in this sample was expected. However, the remaining two strains showed dramatic increase in the number of passages and in the growth rate. A possible explanation for those variations might be reflected in the laboratory workflow used and the high level of the contamination, which can have a significant impact on the DNA collected among the three strains (Drexler and Uphoff, 2002; Ammerman, 2009).

Interestingly, a high number of SNPs was still identified with a high effective rate after mapping them to the reference genome. We demonstrated that despite the high level of similarity in the genomic contents, there was a non- random distribution of the SNPs that were identified per strains. This suggested that there was genetic diversity between the sequences of the three strains (Atkinson *et al.*, 1999; Al-Qassab, Reichel and Ellis, 2010; Regidor-Cerrillo *et al.*, 2013; Calarco, Barratt and Ellis, 2018). These results further support the hypothesis that there is a correlation between the higher polymorphic genes that have non-synonymous SNPs and the pathogenesis of the parasite; this is further supported by the clustering of gene families in some specific regions across genomes that evolve rapidly, specifically in subtelomeric regions (Barry *et al.*, 2003) as confirmed in other pathogenic parasites like *Plasmodium* spp (VAR), *Trypanosoma* Cruzi (VSG) (Berriman *et al.*, 2005; Kyes, Kraemer and Smith, 2007; Jackson *et al.*, 2012; Petter and Duffy, 2015).

In this study, the SNP data analysis and the subsequent cluster findings showed that there was a significant variation in the number of specific gene families (for example the SRS gene family). This finding of patterns of clustering was consistent with previous findings that concluded these genes have a key role in controlling the machinery of immune invasion due to the cellular location on the parasite's surface (Jung, Lee and Grigg, 2004; Risco-Castillo *et al.*, 2011; Reid *et al.*, 2012; Wasmuth *et al.*, 2012; Adomako-Ankomah *et al.*, 2014). These results provide further support for the hypothesis that *NC-Bahia* strain has the highest level of divergence due to the dramatic number of duplication events that have occurred in specific regions, although the majority of proteins were of unknown functions. This finding might reflect, and be in good agreement with, intra-strain differences that were observed previously (Gondim *et al.*, 2001; Regidor-Cerrillo *et al.*, 2013).

# Chapter 5: The use multiple strain sequencing to define variants contributing to phenotypic changes among *T. gondii* isolates

## 5.1 Introduction

*Toxoplasma* diversity studies focused on the three main clonal strains; GT1 (type I), ME49 (type II) and VEG (type III) have revealed that there were highly polymorphic regions that are likely responsible for the unique biological impacts on the *T. gondii* biology. Data from those previous studies also suggest that there was geographical segregation among *T. gondii* strains related to the genetic diversity revealed in the clusters of highly abundant specific groups (Su *et al.*, 2012; Wang *et al.*, 2012; Shwab *et al.*, 2014; Lorenzi *et al.*, 2016). In addition to SNPs, CNVs were found in *T. gondii* when compared to other apicomplexan members with a clear trend of duplications in gaps and on chromosome ends. Notably, a large number of gene sets were significantly enriched and predicted to encode rhoptry proteins (ROPs), dense granule proteins (GRAs), micronemes proteins (MICs) and surface antigen (SRSs) that have been previously identified and known as species - specific genes to influence traits such as host cell attachment, immune evasion, transmission, virulence, host range and phenotypes (Adomako-Ankomah *et al.*, 2014; DeBarry and Kissinger, 2014; Cheng *et al.*, 2015; Lorenzi *et al.*, 2016). Together, these studies provide important insights into the patterns of diversity of those amplified genes that have been reported in different apicomplexan pathogens (Barry *et al.*, 2003; Berriman *et al.*, 2005; Bontell *et al.*, 2009; Wasmuth *et al.*, 2009, 2012; Reid *et al.*, 2012; Kemp, Yamamoto and Soldati-Favre, 2013; Petter and Duffy, 2015; Reid, 2015; Sharif *et al.*, 2017).

One of the greatest challenges that has been noticed in resequencing projects is the quality of unmapped reads that fail to map to the primary reference genome due to many reasons such as the sensitivity of mapping tool used, sequencing errors, contaminations and repeats. In light of recent *de novo* assembly analysis of the sequences of unmappable reads it is becoming extremely difficult to ignore the existence of novel and functional genes between different strains of *T. gondii* across the variable genomes that contain relevant strain-specific sequences per strain.

Despite the large availability of the reference primary sequences in public databases, corrections have been made to improve the gene annotations and understand the evolution and diversity of those strains that would be more useful to predict the full gene complement, which may have covered significant aspects of virulence effectors by improving and deciphering the current genomes of *T. gondii* (Hassan *et al.*, 2012; Whitacre *et al.*, 2015; Van Der Weide *et al.*, 2016; Usman *et al.*, 2017).

## 5.2 Aims of the chapter

This chapter contextualizes the research by providing comparative genomic information by looking to the global genetic diversity that measured the phenotypic variation between distinct *T. gondii* strains. Using genome-wide SNPs and CNVs analysis, the sequences divergence level of 6 sequenced *Toxoplasma* isolates will be analysed to determine the correlation between the genetic divergence and the effects of the SNPs and CNVs based on their locations across the genomic regions. This work had multiple goals.

1. To investigate the variation between strains after generating raw sequence data. After mapping to the currently available reference genome, the pattern of SNPs will be examine and determined in each genome per strain by comparing the SNPs rate in non-conserved and conserved regions between the six *T. gondii* isolates from distinct geographical locations and varied hosts to seek novel SNPs.

2. To get more insight into the significant biological importance of the highly polymorphic candidate genes, which were involved in host - parasite interaction. Gene ontologies will also be assigned to the candidate genes to count the enrichment per gene list and identify overrepresentations of GO terms in general and specific pathways between strains reflecting that there was a significant degree of overall difference in gene functions between different strains of *T. gondii*.

3. To identify novel protein coding sequences that are encoded by reads not mapped to the reference genome. We postulated that there are some novel *T. gondii* genes in different gene families and we hope to discover accurate gene models and gene annotations that are present in the *T. gondii* genome and may have potential function in pathogen-host interactions.

To our knowledge, this will be a detailed further comparative genomic investigation of six different strains that differ dramatically in their single nucleotide polymorphism and copy number of variations. Hence, this project aimed to address the following research questions: Does the variation reflect geographical distributions? Why do the different isolates show different patterns of diversity? What is the association between the genes families and virulence in *T. gondii*? Are there positive correlations between the copy number of variations and clustering in specific locations across the genome? And finally, what are the uncommon features of duplicated genes or genomic segment that are not seen in *N. caninum* strains?

## 5.3 Results

### 5.3.1 Data generation and sequenced read alignment of the *T. gondii* isolates to the entire *T. gondii* ME49 reference genome

We generated whole genome sequencing data from six strains of *T. gondii (T. GT1, T. MAS, T. CAST, T. VEG, T. P89* and *T. COUG*) (Table 5.1). From the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/?term=SRR6793863), we downloaded the reads from *T. gondii* strain ME49. Before further data processing, all the sequencing row reads were trimmed to remove low quality reads. The sequence contamination representing genomic DNA from the mammalian tissue culture were removed successfully. We included all the reads that passed all the filtration criteria in our comparison and exported to the next downstream analysis including mapping to the reference genome of the *T. gondii* strain ME49, SNP detection and *de novo* assembly for the unmapped reads (Table 5.2).

After cleaning reads, all the reads that were most likely to contain specific sequences per strain were mapped to the known complete reference genome of *T. gondii* strain ME49. It has been found that there were large discrepancies in the proportion of the reads obtained per sample after alignment as shown in Table 5.2. Due to the data being generated, the percentage of mapping was significantly different per strain. We would expect different mapping percentages within isolates based on the number of bases mapped per chromosome to the primary reference genome and it might provide a first indicate of genetic diversity between *T. gondii* strains due to the poorest mapping compared to the highest mapped strains. As to the mapped reads, it can be seen from our data that the *T. COUG, T. MAS* and *T. CAST* which are known as non-reference strains had significantly lower mapping percentages than the other isolates due to the high number of sequence reads that remained unmapped with 35.77, 20.9 and 16.1 % respectively. From our mapping data, we can see that *T. GT1* resulted in the highest rate of mapping compared to the remaining five *T. gondii* strains. (see Figure 5.1). By examining the coverage and the mean converge per chromosome and per strain, from our data, there was no specific bias between the 14 chromosomes in terms of the depth of reads due to the similar coverage percentages for all chromosomes for each strain with the exception of *T.GT1* for chromosome IX (Figure 5.2).

From the Figure 5.2 below, we can see that there were significant different coverage profiles between the six strains of *T. gondii.* The coverage was actually higher in *T. CAST, T. MAS* and *T. COUG* strains than the remaining three isolates reflected in the count of bases being covered per chromosome. More specifically, the coverage per chromosome differed between the six isolates from chromosome to chromosome and from strain to strain as plotted in Figure 5.2. The minimum coverage was noticed in references strains *T. VEG, T. GT1* and in non-reference strain *T. P89* with a clear trend of decreasing in the mean coverage as well. As a result, the genome coverage obtained for the *T. gondii* isolates was sufficient to be applied for variant detection (SNPs) between the representative isolates.

**Table 5.1:** Summary of sequence read data generated by paired-end sequencing with 150 bp average length of reads of six samples of *T. gondii* before and after adapter removal and low Phred score trimming.

| Sample | Total number of raw reads | Total number of trimmed reads | R1 forward | R2 reverse | R0 |
|---|---|---|---|---|---|
| ***T. GT1*** | 54062474 | 53392164 | 26544847 | 26544847 | 302470 |
| ***T. MAS*** | 60830968 | 59882781 | 29725376 | 29725376 | 432029 |
| ***T. P89*** | 76937570 | 75169813 | 37329170 | 37329170 | 511473 |
| ***T. CAST*** | 65355958 | 64635070 | 32122381 | 32122381 | 390308 |
| ***T.VEG*** | 71433580 | 70662968 | 35119089 | 35119089 | 424790 |
| ***T. COUG*** | 57908686 | 56839057 | 28222506 | 28222506 | 394045 |

**Table 5.2:** The summary statistics of mapping analysis of six isolates of *T. gondii* sequences that mapped to the known reference genome of *T. gondii* strain ME49 version 29 downloaded from (http://toxodb.org/common/downloads/Current_Release/TgondiiME49/fasta/data/).

| Data type | *T.GT1* | *T.MAS* | *T. P89* | *T. CAST* | *T.VEG* | *T. COUG* | *T. ME49* |
|---|---|---|---|---|---|---|---|
| **Reference size** | 65,669,794 | 65,669,794 | 65,669,794 | 65,669,794 | 65,669,794 | 65,669,794 | 65,669,794 |
| **Total number of trimmed reads before removing contamination** | 53,3921,64 | 60,830,968 | 75,1698,13 | 65,3559,58 | 71,4335,80 | 57,9086,86 | - |
| **Total number of clean reads after removing contamination** | 28,006,138 (52.4%) | 59,685,327 (98.1%) | 18,019,142 (23.9%) | 64,475,885 (98.6%) | 33,185,533 (53.3%) | 56,476,149 (97.6%) | 86,345,805 |
| **Total number of unclean reads contain contamination** | 25,386,026 (47.5%) | 1,145,641 (1.9%) | 57,150,671 (76%) | 880,073 (1.4%) | 38,248,047 (53.5%) | 1,332,537 (2.3%) | - |
| **Total number of mapped reads to the reference genome** | 27,700,555 98.91% | 47,573,527 79.8% | 15,156,348 84.11% | 54,064,958 83.85% | 30458593 91% | 36,622,468 64.73% | 72,795,732 84.31% |
| **Total number of unmapped reads after mapping to the reference genome** | 305,583 1% | 12,111,800 20.29% | 2,862,794 15.89% | 10,410,927 16.15% | 2,726,490 8.22% | 19,953,681 35.27% | 13,550,073 15.69% |
| **Genome Coverage (X)** | 63.97X | 136X | 41.15X | 149X | 75.8X | 129X | 197X |
| **Mean Coverage** | 60 | 103 | 25.56 | 118.5 | 64.16 | 79.9 | 161.7 |

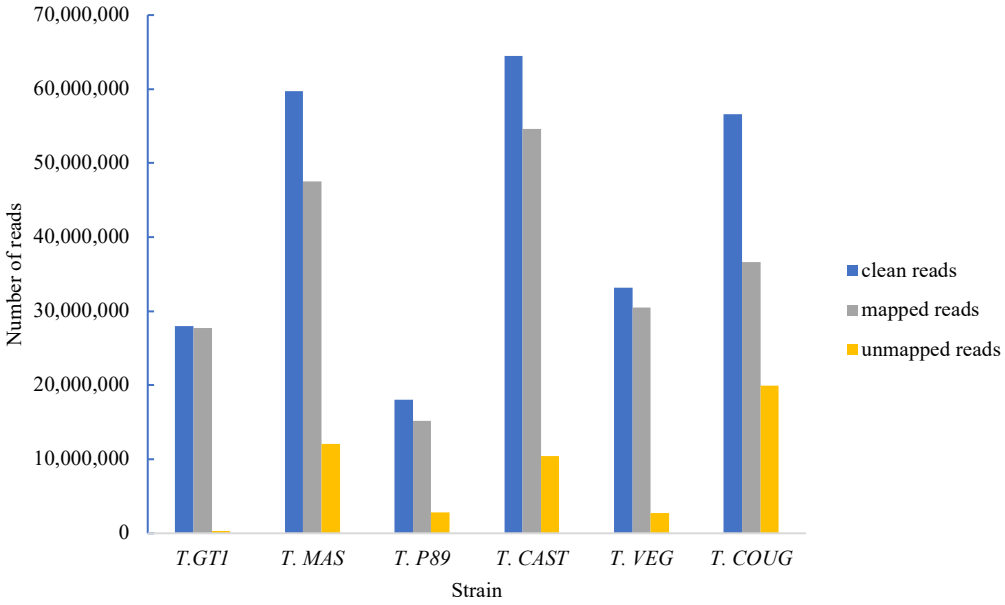



**Figure 5.1** Overview of type of read counts obtained from six strains of *T. gondii* when mapped to the *T. gondii* reference genome. The highest percentages of clean reads after removing contamination was in *T. CAST, T. MAS* and *T. COUG*.

### 5.3.2 SNP analysis

#### 5.3.2.1 Identification of SNPs between the six strains of *T. gondii*

Following mapping of reads to the reference genome assembly, SNPs were called using the GATK3 tool. A total of 494,877, 397,610, 539,562, 529,126 , 342,568, 367,175 and 4728 putative SNPs were identified over the entire genomes of *T. GT1, T. MAS, T. P89, T. CAST, T. VEG, T. COUG* and *T. ME49* respectively (Table 5.3). The SNP rates differed per strain. Collectively, the number of shared SNPs were 546 SNPs within the strains with highest level of conservation noticed in *T. ME49* as we expected. Further analysis was performed to define a degree of uniqueness and overlapping within strains based on the number of SNPs that were only identified in each strain as mentioned in Chapter 2. From the Table 5.3 below, the breakdown of the percentages of the degree of the uniqueness and shared SNPs per strain were observed.

We found that there was inter-strain variation based on the degree of uniqueness that revealed there was a high degree of uniqueness between strains. This suggests there was a genetic diversity impact between strains that play an important function in the phenotypic variabilities. The strain with the highest percentage of uniqueness *T. COUG* strain with 159029 SNPs. As we described earlier in Chapter 1, the reference genome of *T. gondii* strain *ME49* was sequenced by capillary sequencing methods corresponding to a genome depth of 26.5X. The total number of SNPs in this reference genome of *T. gondii* ME49 was estimated based on local variations in the different genomic locations per chromosome. By comparing the estimated number of SNPs in *T. gondii* strain ME49 to the reference genome we found 228,3933 SNPs; this provided evidence that the strains had novel SNPs resulting from intra-lineage divergence between different strains of *T. gondii*. Despite the lower coverage identified in T. *P89, T. GT1* and *T. VEG* strains, a much greater proportion of SNPs was detected in *T. P89* and *T. GT*. Figure 4.4 presents the coverage per chromosome of the *T. gondii* strains, more specifically, in *T.GT1* that showed high coverage in chromosome IX that will be examined later for a sensitive look to specific regions in this chromosome.

**Table 5.3**: Summary of SNP calling generated by GATK3 tool. In this table, SNPs were generated after mapping to the *T. gondii* reference genome available from (www.ToxoDB.org).This was done using VCF- Compare after filtering the data using final filtered draft of VCF file per strain.

| Strain | SNPs unique to the strain | SNPs Shared with another 5 strains | SNPs Shared between all 6 isolates | Total SNPs |
|---|---|---|---|---|
| ***T. GT1*** | 59,901 | 434,430 | 546 (0.1%) | 494,877 |
| ***T. MAS*** | 125,721 | 271,343 | 546 (0.1%) | 397,610 |
| ***T. P89*** | 79,427 | 459,589 | 546 (0.1%) | 539,562 |
| ***T. CAST*** | 66,422 | 462,158 | 546 (0.1%) | 529,126 |
| ***T. VEG*** | 30,071 | 311,951 | 546 (0.2%) | 342,568 |
| ***T. COUG*** | 1,59029 | 20,7600 | 546(0.1%) | 367,175 |
| ***T. ME49*** | 1,967 | 2,215 | 546(11.6%) | 4728 |

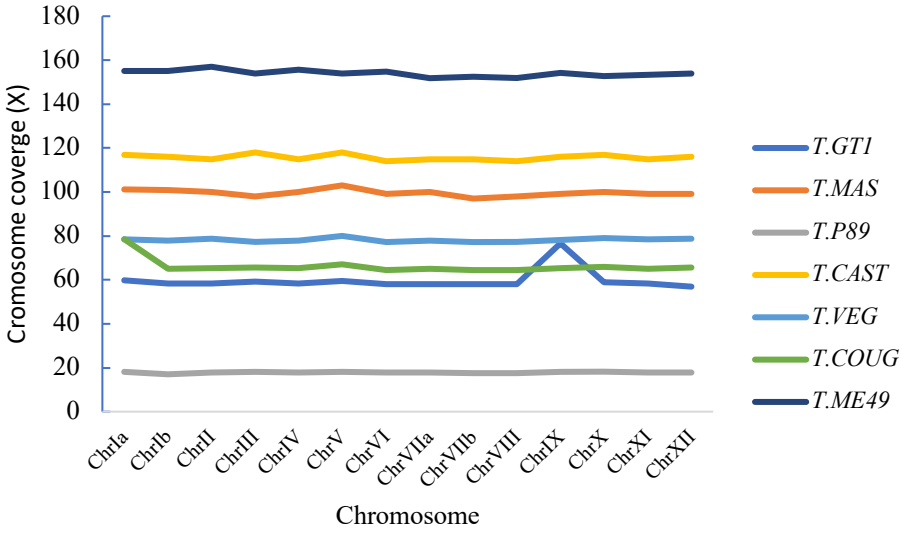



**Figure 5.2:** The summary statistics of coverage per chromosomes of the *T. gondii* strains. As legend shown *T.GT1* showed high coverage in chromosome IX. The y axis shows the chromosome coverage (x) and the x axis donated to the chromosome.

**Table 5.4:** The number of SNPs estimated in all *T. gondii* strains were called by using GATK3 that identified difference from the reference genome of *T. gondii* strain *ME49*. Here, it has been shown that there was a diversity per strain and per geographical origin. The highest count of SNPs represented was in *T. P89* strain and the lowest count was identified in *T.VEG* strain.

| **Data type** | ***T. GT1*** | ***T. MAS*** | ***T. P89*** | ***T. CAST*** | ***T. VEG*** | ***T. COUG*** |
|---|---|---|---|---|---|---|
| Total number of SNPs | 494,877 | 397,610 | 539,562 | 529,126 | 342,568 | 367,175 |
| % SNPs in coding DNA sequence (CDS) | 9.2% | 9.3% | 9.8% | 9.2% | 9.8% | 9.1% |
| Number of nonsynonymous SNPs | 65,713 | 52,953 | 71,593 | 70,344 | 45,193 | 49,278 |
| Number of synonymous SNPs | 46,182 | 37,563 | 50,123 | 49,316 | 31,070 | 33,110 |

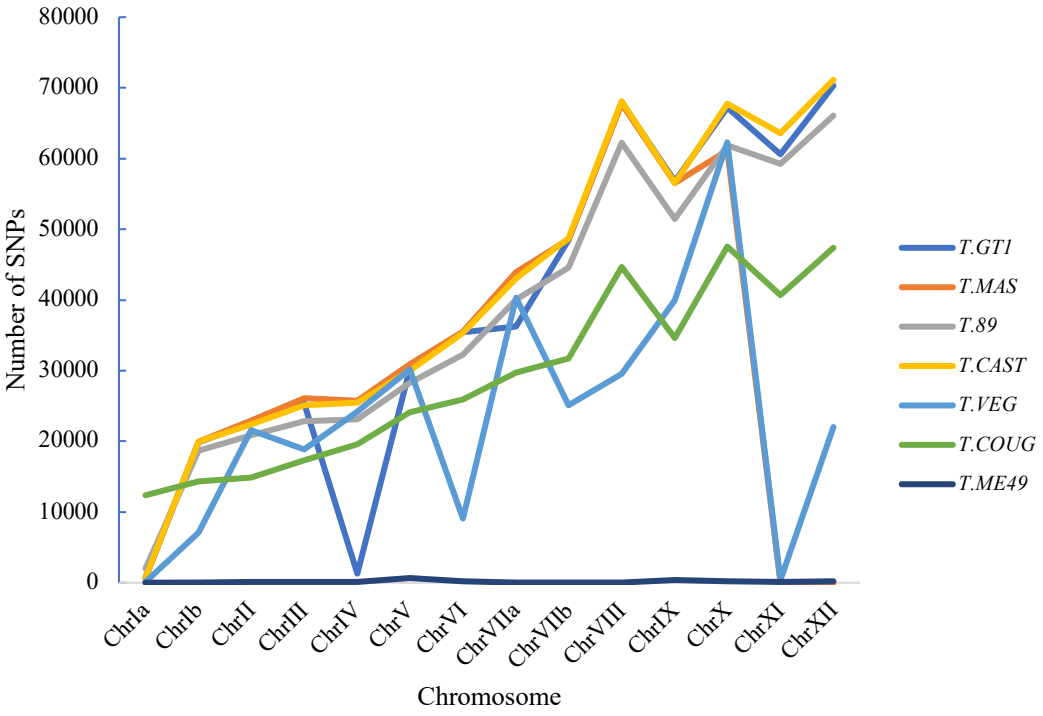



**Figure 5.3:** The number of SNPs detected in each of the *Toxoplasma* isolates (*T. GT1, T. MAS, T. P89, T. CAST, T.VEG, T. COUG* and *T. ME49*) based on the location in different regions of the genomes. The highest number was noticed in *T. P89* (grey) and *T. CAST* (orange) strains and the lowest count was in *T. ME49* strain.

### 5.3.2.2 Distribution and density of SNPs in the *T. gondii* isolates

The distribution of the SNPs identified per strain with their corresponding base changes including the transversion (Tv) /transition (Ts) ratio in the six isolates of *T. gondii* were plotted in Figure 5.4. The SNPs were plotted based on the different genotypic changes they caused per strain. The transversion events including; A > C, C < A, G >T, T > G and A > T and transitions including; C >T, T > C, G >A and A > G were generally identical. In all the six strains of *T. gondii*, we found that transitions were more common than transversion events which was expected in term of occurrence and less likely to produce a difference in the amino acid sequence compared to transversions that result in a silent mutation known as low impact. Further comparison was performed to compare the Ts /Tv ratios of all of the SNPs between the six strains, highlighting that there was no significant difference in the Ts /Tv ratios that approximately equal between them with a slightly increase in the Ts /Tv ratio in *T. COUG* strain compared to the other five *T. gondii* strains. The lowest ratios were detected in *T.VEG* and *T. P89* strains due to low coverage obtained after sequencing.

To determine the density per chromosome as shown in Figure 5.5, the SNPs density per 1000/kb for each chromosome and strain was calculated. Most of the SNPs were distributed randomly across the 14 chromosomes. However, high density of SNPs were primarily found in sub telomeric regions in nine chromosomes including II, III, IV, V, IX, VIIa, VIIb, IX and X. A relatively higher density of SNPs was also noticed in central regions of chromosomes Ib, VI and VIII. The total number of SNPs per chromosome is shown in Figure 5.6, we observed that the chromosome Ia showed the smallest cluster of SNPs reflecting the lowest diversity identified compared with other chromosomes in the six *T. gondii* isolates, as we expected due to the shortest length of this chromosome.

High diversity was noticed in Ib, II, IX, VI, VIIa, VIIb, VIII, X and XII that confirmed dramatic changes in the density of the SNPs. It has been found that there was pattern of telomeric clustering of the SNPs that were located in chromosomes II ,VIIa, VIIb, X, XI and XII, which suggested that these chromosomes might contain protein families with a key role in host parasite interactions and also influence the genetic diversity between different strains. The overall genomic pattern of clusters of high and low diversity regions revealed that the density of SNPs was positively correlated with specific regions in the genomes that have polymorphic genes across different loci.

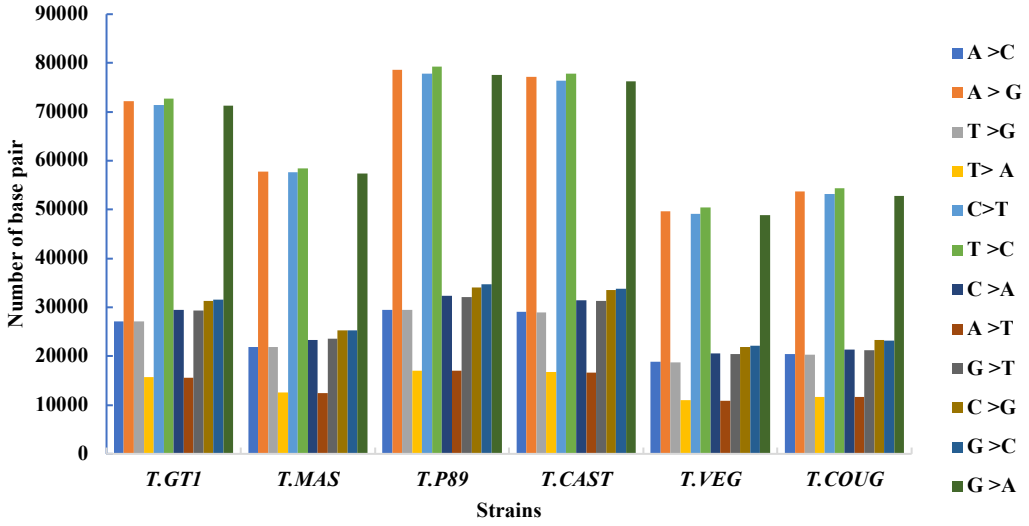



**Figure 5.4:** The number of base pairs per strain and the corresponding changes they caused. This shows that despite variations in the number of SNPs called in whole data sets for the six samples, the transition/transversion ratio was nearly identical, except in *T. COUG* strain.

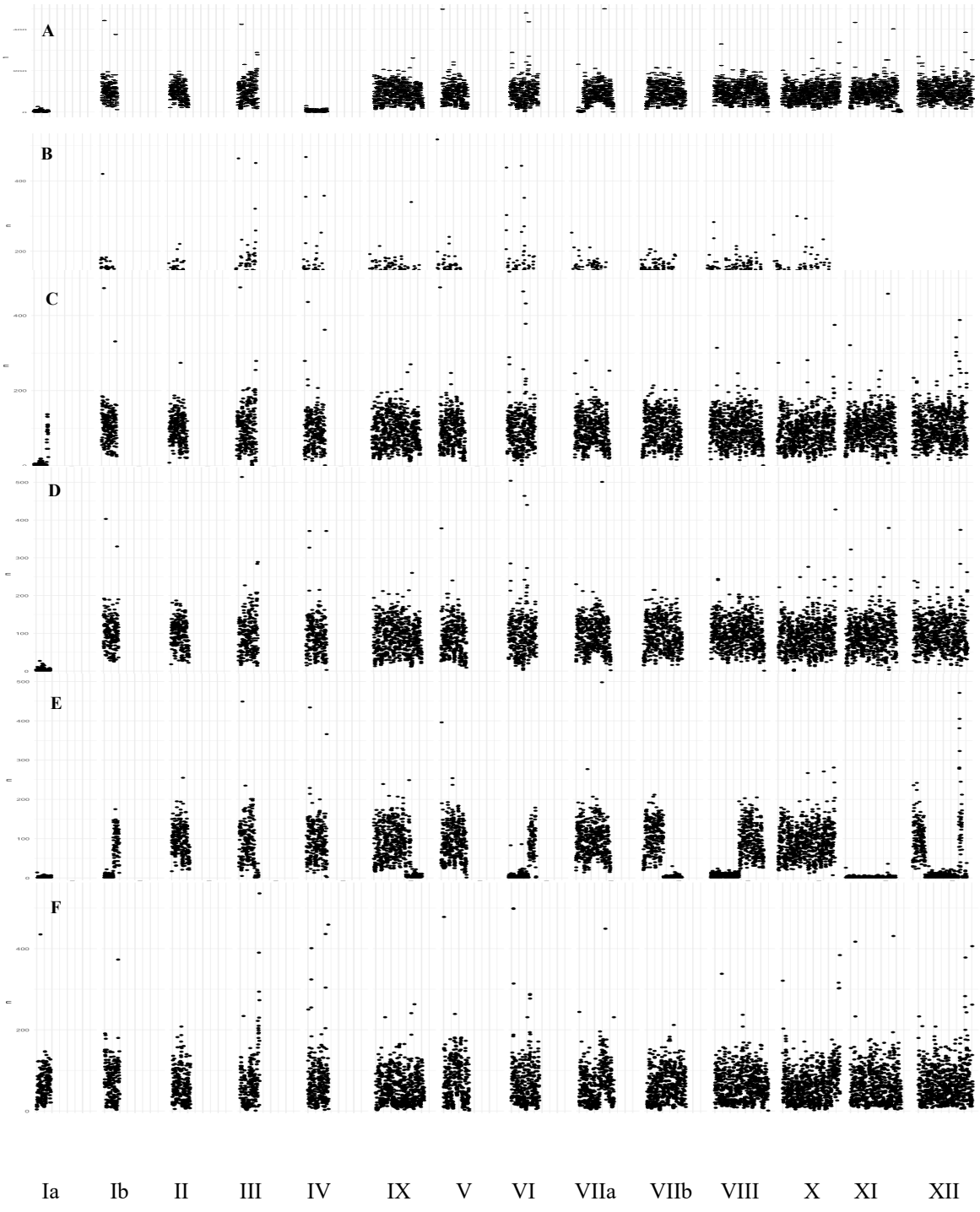



**Figure 5.5**: The SNP density per 1000/kb in *T. gondii* strains across the 14 chromosomes. **A)** *T. GT1*, **B)** *T. MAS*, **C)** *T. P89*, **D)** *T. CAST*, **E)** *T. VEG*, **F)** *T. COUG*.

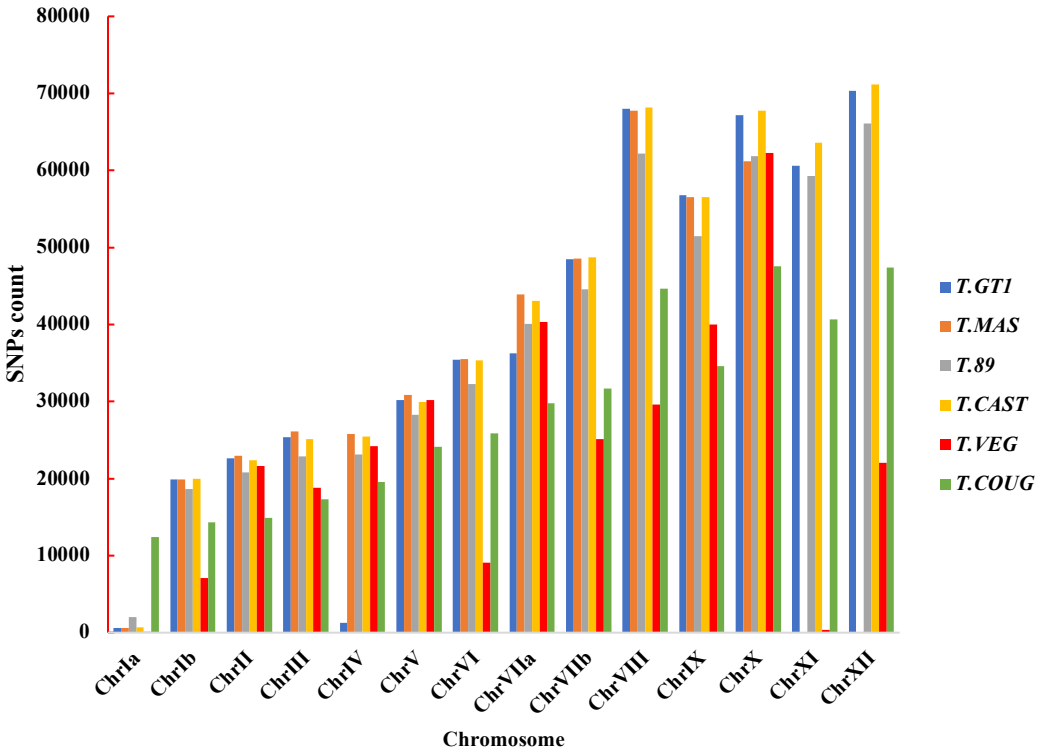



**Figure 5.6:** The total number of SNPs per chromosome in the six *T. gondii* strains aligned to the reference genome of *T. gondii* strain ME49; the SNPs distribution over the 14 chromosomes are shown in different colours. Blue colour denoted *T. GT1*, orange *T. MAS*, grey *T. P89*, yellow *T. CAST*, red *T. VEG* and green *T. COUG* coulure. The highest number of SNPs was observed in six chromosomes II, IX, VI, VIII, X and XII. The lowest count of SNPs identified in chromosome Ia in all the *T. gondii* strains.

### 5.3.3 Investigation of genomic diversity within different sequence classes

The analysis of functional variants between the reference genome of *T. gondii* strain ME49 and the six re-sequenced strains was performed using the SnpEff software. This allowed us to analyse which regions of individual chromosomes were highly enriched and diverse and whether those SNPs were clustered in specific regions or randomly distributed across the genomes. To predict their effects, the SNPs were analysed according to genomic region (coding and non-coding) and likely impacts. All the annotated SNPs were investigated to evaluate in which region the SNPs was found. The varied effects were identified and compared per region as shown in Figure 5.7. It can be seen that by far the greatest number of SNPs were observed in downstream and upstream of coding regions with 30% and 28.5% of the total effects predicted across all *T. gondii* strains, respectively. This pattern of variability might have an impact on the gene expression and regulatory role of particular genes. Our results showed that the second greatest number of SNPs were predicted to have effects in intronic and intergenic regions with 15% and 10 % respectively. As we expected, limited polymorphism was found in the coding regions (exons) among *T. gondii* strains with an average 9.4 % as shown in Table 5.4 above. In the *T. P89* and *T. CAST* strains, it has been found that the percentage of variants in the coding (CDS) or exonic regions was higher compared to other *T. gondii* strains (see Figure 5.7). Conversely, this was lower in the *T. VEG* strain.

The effects were also predicted in intergenic regions in the six strains with high percentage of effects in the *T. CAST* strain that accounted for 20% of the total effects in all strains. Together, the present findings confirm significant variations of selection pressures per region. The SNPs annotated were also assigned as synonymous (silent) and non-synonymous (missense and nonsense) substitutions in all strains. This was done to reveal which mutations were more frequent in the genes that have mutations. More significantly, in the specific genes that control the host-parasite interaction and examine the role of polymorphism in genetic diversity and phenotypic changes between strains. For comparison, there was significant variation in the number of SNPs per types that grouped into non-coding and coding SNPs. The vast majority of the SNPs were non-coding substitutions as described above. However, most of the coding SNPs (58%) were annotated as nonsynonymous SNPs.

We found that 43.7% of SNPs represented synonymous SNPs mapping to coding regions, while only 0.5% were assigned as nonsense and had a missense / silent ratio greater than 1.4 of the total SNPs types predicted in all six isolates (Figures 5.8 & 5.9). Collectively, the above analysis, SNPs classified into different types revealed strong evidence of divergence in the proportions of nonsynonymous (pNS) and synonymous (pS) and the ratio of pNS/ps SNPs between the linages of *T. gondii*; the non-synonymous mutations were more frequent than the synonymous SNPs due to changes in the coding sequences per strain as we expected. Hence these data suggest that the differences in patterns of polymorphism might be directly implicated in pathogen-host interaction in different genomes of *T. gondii* isolates.

Further SNPs analysis was performed to determine the putative impact of the SNPs which could be categorized into four impact groups, low, moderate, modifier and high impact. As we can see from Figure 5.10, by comparing the overall impacts of all the mutations, our results demonstrated that the vast majority of the SNPs were predicted to cause modifying impact. The modifier SNPs were considered as noncoding SNPs that mainly affected noncoding genes in both UTRs, intergenic SNPs and in putative regulatory sequences. Moderate impact was mainly observed in the proportion of the nonsynonymous (missense) SNPs that causing a harmless effect on protein production. This was followed by low impacts with no greater than 5 % out of the total impacts, which were silent (synonymous) SNPs that functionally do not change the predicted protein sequence. High impact SNPs were significantly lower, consisting of no more than 0.2 % of the total number and thus were rare compared to the other three impacts. The details of the sub classification of the impacts group (moderate, modified and low) was plotted in Figure 5.11 per strain. However, the high impact SNPs were also investigated to identify the polymorphic genes (see next section 4.3.5.4). In addition to this, all the related pathways were examined to assess whether there was a significant association between the particular pathways for each impact with the GO terms enriched. This allowed us to see what the genes were that are assigned to specific groups of gene families per each impact and per strain.

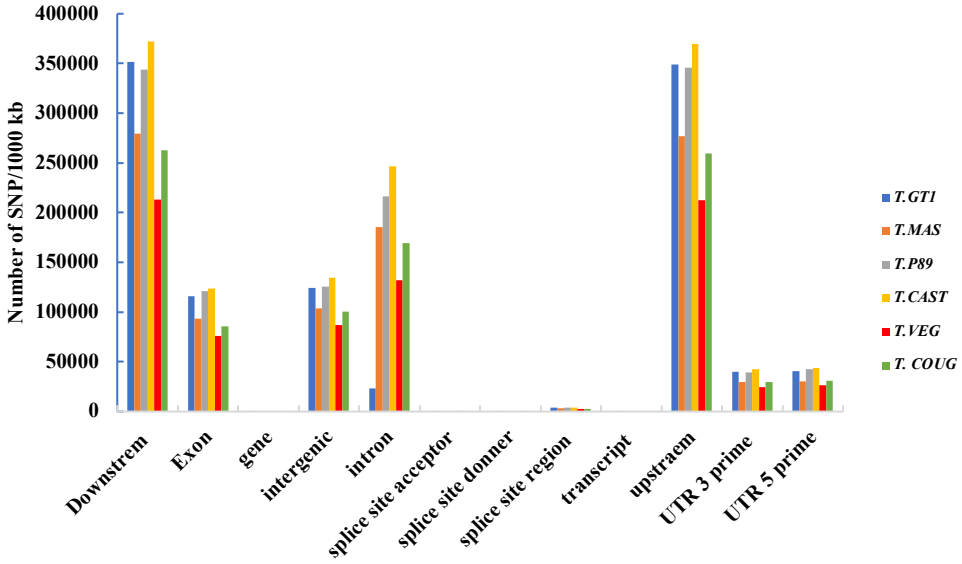



**Figure 5.7:** The number of SNP effects by genomic region. SNPs were grouped into different regions based on their locations in the annotated genomes of *T. gondii*. These include; introns, intergenic regions and untranslated regions; 5' UTR (leader) and 3' UTR (trailer), exons and splice sites. The vast majority of the SNPs were predicted to have effect on the downstream and upstream regions across all the *T. gondii* strains.

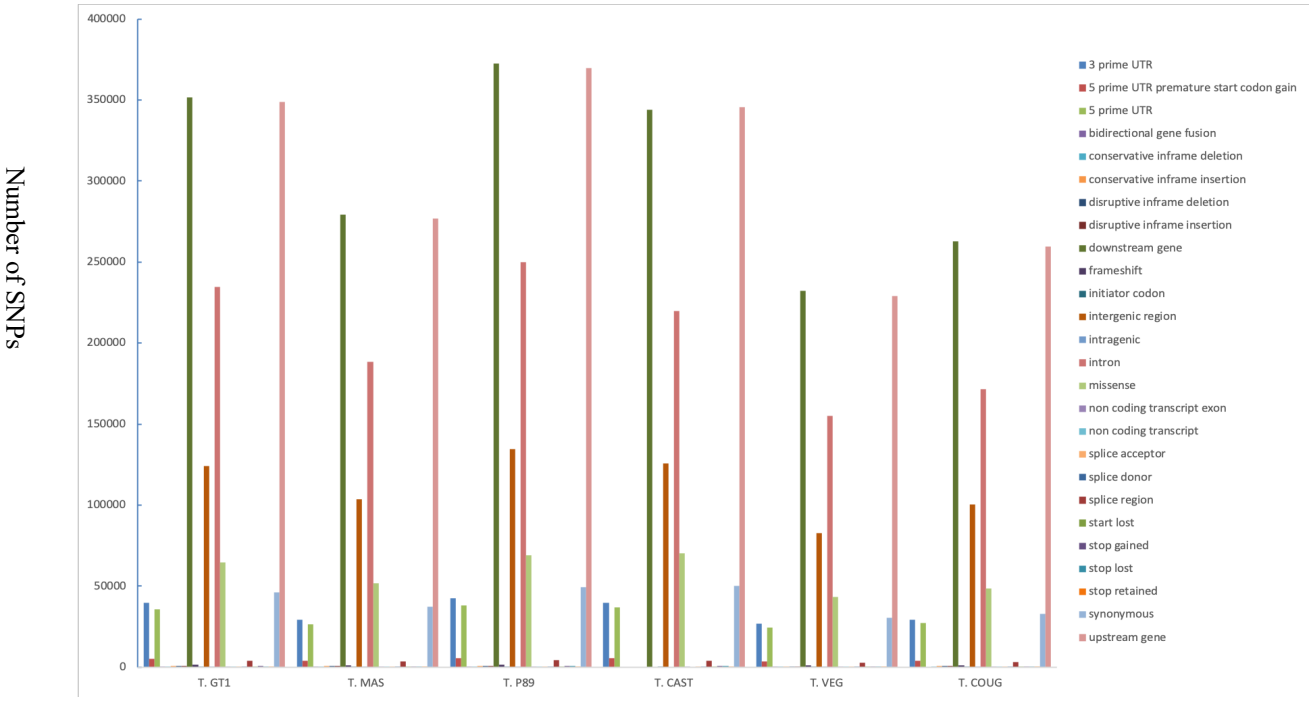



**Figure 5.8:** The number of SNPs annotated that grouped into different types in the six mutated genomes. The majority of SNPs were from downstream in dark green and upstream in pink across all the *T. gondii* strains.

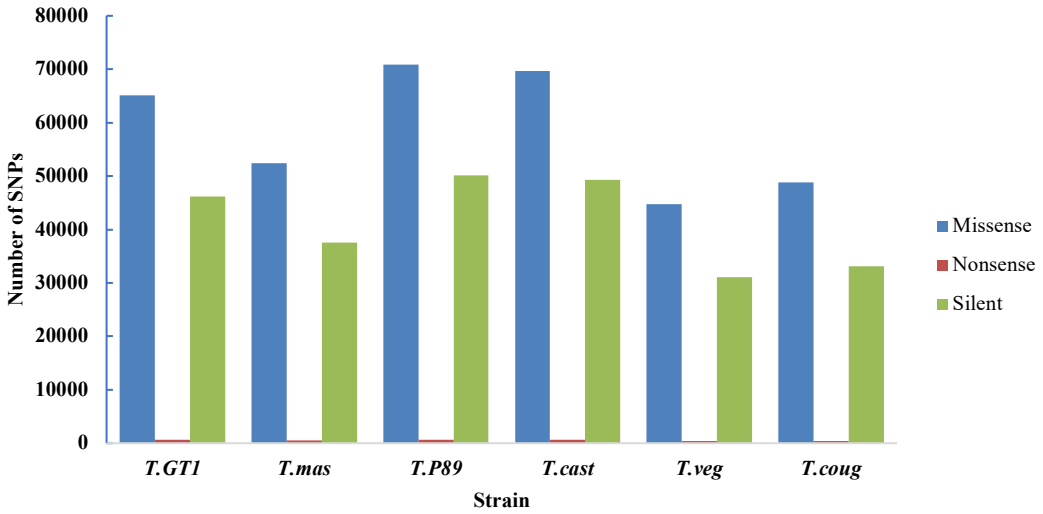



**Figure 5.9:** The number of SNPs annotated that grouped into different functional effects in the six mutated genomes. The majority of SNPs were coding SNPs that annotated as missense SNPs (non-synonymous) blue and nonsense showed in red. Silent (synonymous) SNPs were showed in green across all the *T. gondii* strains.

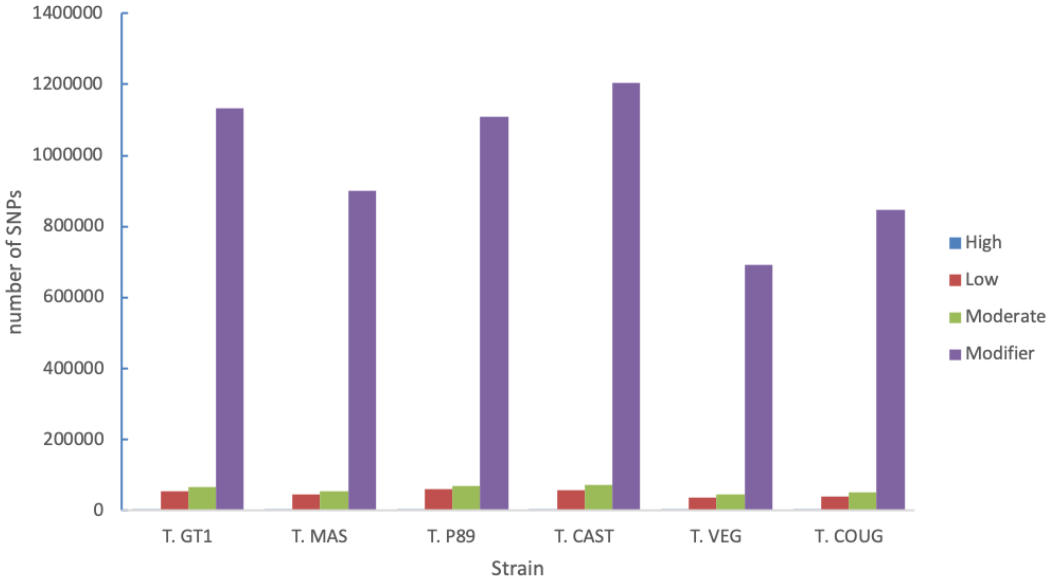



**Figure 5.10:** The number of SNP effects by impacts. The SNP effects were grouped into modifier, moderate, low and high. The highest percentage of impact was observed for modifier that were mainly located in upstream, downstream, intergenic and UTR regions in all the strains of *T. gondii*.

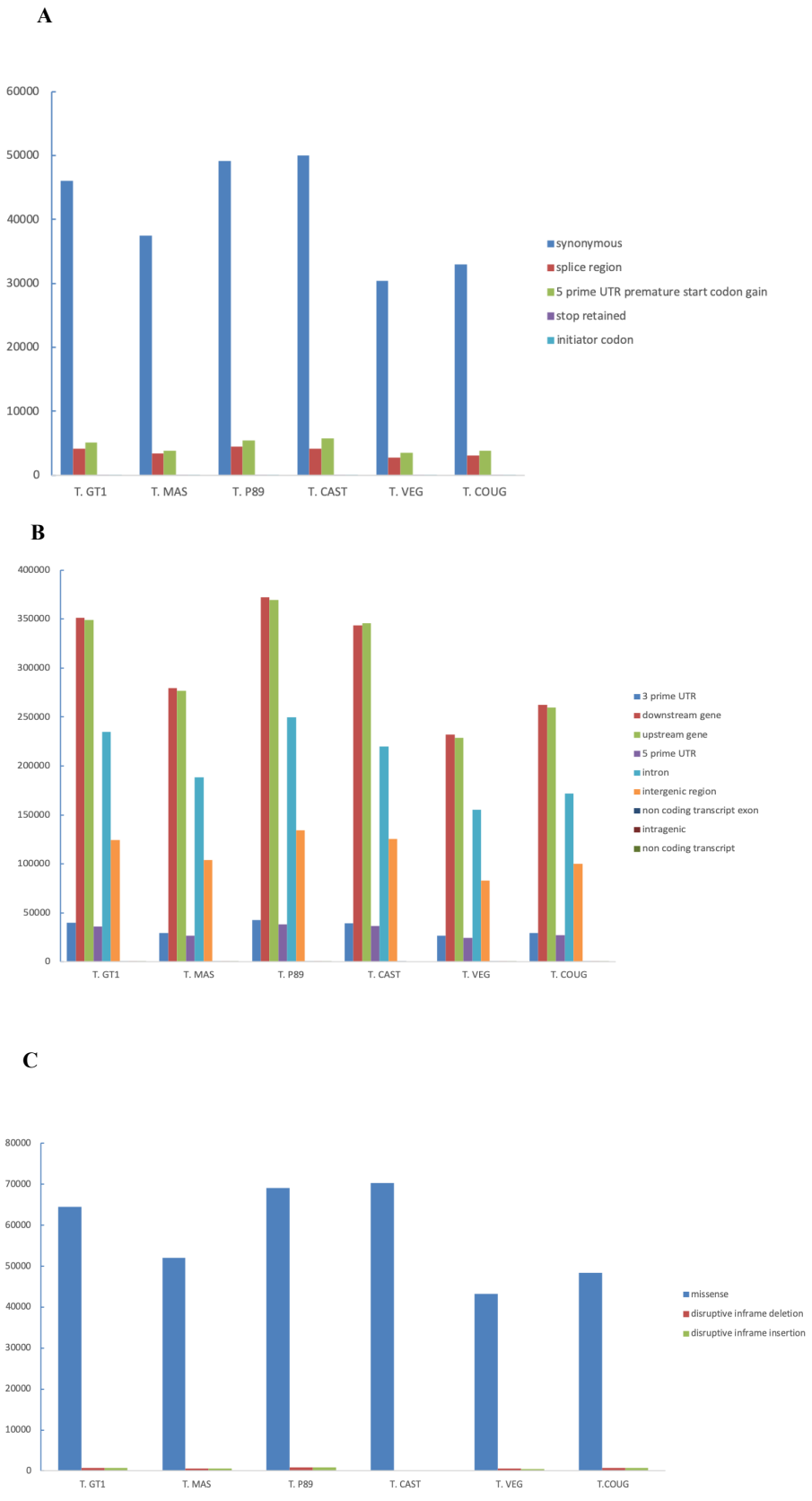



**Figure 5.11:** The percentage contribution of sub-classification of the four impacts; A) Low, B) Modifier and C) Moderate across all the *T. gondii* strains**.** SNP effects were categorized by each effect per impact as low (synonymous coding), and modifier (upstream, downstream, intergenic, UTR) and moderate (missense) that mainly were nonsynonymous.

### 5.3.4 Investigating the of frequency of SNPs in the most diverse genes within each strain

In the previous section, the genetic diversity was described between *T. gondii* strains per region, type and impact based on the structural annotation of the genome. To obtain the genes associated with SNPs in the genomes, the total number of SNPs affected per gene were counted to determine if those genes were under selective pressure or not. The number of SNPs per strain and per chromosome was varied as we confirmed earlier in section 5.3.3.3, (see Figure 5.5 and 5.6). We found that 8378, 6313, 8320, 8390, 7058 and 8429 annotated genes contained one or more SNPs, with an average frequency of 3 SNPs /gene across all six strains.

As might be expected, the number of genes that have nonsynonymous SNPs was considerably higher than the number of genes harbouring synonymous SNPS. The results of this analysis revealed that 65102, 52470, 70923, 69,679, 44777 and 48830 SNPs within 6451, 5136, 6818, 6762, 4471 and 6247 genes showed nonsynonymous (missense) substitutions. Furthermore, the nonsynonymous SNPs that potentially gave high impact (nonsense) substitutions were located in 1496, 1146, 1003,1567, 629 and 1160 genes in *T. GT1, T.MAS, T. P89, T. CAST, T. VEG* and *T. COUG* strains, respectively. We detected 46182, 37563, 50,123, 49,361, 31,070 and 33110 SNPs in 6759, 5020, 6679, 6619, 4238 and 5876 genes resulting in synonymous substitutions (silent) in *T. GT1, T. MAS, T. P89, T. CAST, T. VEG* and *T. COUG* strains respectively (see Figure 5.9). The above analysis was based upon the SNP proportions in the coding regions and revealed that there was a significant difference in the number of SNPs in coding regions, thus highlighting sequences that might be associated with traits of interest, such as; transmission strategies, virulence and improving the fitness of the *T. gondii* in different environment.

As we stated above, the coding SNPs were assigned into synonymous and nonsynonymous based on the current gene annotations of the reference genome *T. gondii* strain ME49. In all the six strains of *T. gondii*, the largest number of genes with the highest number of SNPs were annotated as hypothetical proteins with unknown functions as we expected. Furthermore, there was a significant increase in the frequencies of the nonsynonymous SNPs in regions encoding proteins that including biologically relevant gene families that have an important function.

The rhoptry kinase proteins (ROPs), micronemes (MICs), dense granules (GRAs), surface antigens -SAG named (SRSs), *Toxoplasma* specific families (TgFAMs) including; A, B, C, D and E and Lysine- Arginine rich Unidentified Function (KRUFs) families were expanded in *T. gondii* strains. The frequencies of SNPs in the gene families varied per strain (Figure 5.12). We expanded our analysis to identify evidence of selective evolutionary pressure and compare the divergence of encoded proteins (i.e. SRSs, ROPs, GRAs, MICs and TgFAMs). The accumulations of polymorphisms in the coding regions using the frequencies of nonsynonymous (NS) and synonymous (Sy) between the different representative genomes of *T. gondii* strains might be guiding us to understand the potential impact of the SNPs on the phenotypic changes and genetic diversity that correlate with virulence.

Table 5.5 highlights the different gene families that were uniquely enriched and expanded between the different strains. The SRSs genes appear to have most polymorphic genes among the five gene families with a total number of SNPs of 1012,765,1313, 1238,779 and1264 in 83, 64, 102, 101 and 61genes in *T. GT1, T.MAS, T. P89, T. CAST, T. VEG* and *T. COUG*, respectively. By comparing the polymorphisms between strains, we found that the vast majority of SNPs were in the SAG-related sequence SRS16B gene known as SRS9 (TGME49_320190) on chromosome IV. The highest proportion of nonsynonymous SNPs was in *T. CAST* with 83 SNPs then *T. MAS* and *T. COUG* strains with 78 and 76 SNPs respectively. In *T. VEG* and *T. P89* strains, the same cumulative value of 68 SNPs was detected. This indicated that there was local strain variation between isolates compared to the reference genome of *T. ME49* strain. In the *T. GT1* strain, there was a significant decrease in the number of SNPs with only 47 SNPs in SAG-related sequence SRS48E (TGME49_296640) in chromosome X.

The second family of polymorphic genes that demonstrated substantial variability in the six strains of *T. gondii* was the TgFAMs, including the five subfamilies A - E. From our analysis, we noticed that the TgFAMA family contained more genes than other gene families. In the *T. P89* and *T. COUG* strains, the most variable gene (TGME49_278090) had the highest number of SNPs, 56 and 57 SNPs respectively on chromosome XI. Furthermore, variations also have been identified within this family of genes with 83, 59, 51 and 48 SNPs in *T. VEG*, *T. CAST, T. GT1* and *T. MAS* strains respectively located in the (TGME49_266340) gene.

A third gene family, the ROP genes, comprises 62 genes in the current release of the reference genome of the strain ME49. We detected SNPs within all 62 ROP genes in all six *T. gondii* strains. A subset of the 62 ROP genes had a very high proportion of SNPs, indicating that they might be local strain divergence and the novel SNPs might be in the major virulence factors within the different strains. Interestingly, our data showed a greater polymorphism in some genes that had a significantly higher level of nonsynonymous to synonymous substitutions across the virulent and non-virulent lineages of *T. gondii*. The number of SNPs was dramatically increased in T. P89 with 531 SNPs within 50 genes as shown in Table 5.5.

These findings based on the SNP frequencies revealed a significant number of SNPs in specific locations across the genomes. Most of the SNPs were distributed throughout the key virulence effectors genes located on chromosome VIIb, including ROP 16, ROP17 and ROP 39. In addition, we found that the ROP16 gene had the highest number of SNPs of all the *T. gondii* strains, more specifically in *T. VEG* strain 66 SNPs. The next rhoptry kinase protein that had a higher frequency of SNPs was the ROP39 gene with more than 30 SNPs. The genetically diverged strains harboured different number of SNPs in ROP18 (TGME49_205250) and ROP5 (TGME49_308090), both located on chromosomes VIIa and XII, respectively. The highest rate of polymorphism was detected in *T. P89* with totalling 66 SNPs at ROP18 gene. Additionally, the pattern of polymorphisms was also noticed in *T.VEG* with 49 SNPs in virulent alleles of the ROP5 gene. These data suggest that the extensive variations between alleles at ROP5 and ROP18 proteins will primarily shape the structure of *T. gondii* populations, resulting in significant evidence of positive selection and evolution in different *T. gondii* strain types.

Members of the dense granules gene family (GRAs), however, demonstrated a significant reduction in the number of variations within this family compared to SRSs, TgFAMs and ROPs families. This family comprised 18 members in the reference genome and is associated with the parasitophorous vacuole membrane (PVM) during acute infection and subversion of host immunity. Several SNPs were identified in dense granule proteins. The most variable gene, the dense granule protein (GRA3) located on chromosome X, had 19 SNPs in *T.CAST*. In addition, GRA4, GRA7, GRA8, GRA11were noticed in all the strains.

A total of 26 micronemes (MICs) genes were identified by performing multiple queries from ToxoDB of the reference genome ME49. By comparing the number of SNPs between strains of *T. gondii*, we observed that there were particular variable genes with a much greater number of SNPs in *T. P.89* and *T. GT1* strains, with 151 and 136 SNPs within 23 and 21 genes, respectively. Indeed, one of the most diverse genes amongst the strains was the MIC15 gene shared between three strains of *T. gondii*; T. P89 with 61 SNPS, and a total of 57 SNPs in *T.GT1* and *T. CAST* strains, respectively. The KRUF gene family was expanded in the *T. gondii* strains compared to other gene families described above. This family was found in all strains of *T. gondii*. All the variations within polymorphic gene families were summarized in Table 5.5 below, and they are directly implicated in host parasite interactions. Overall, the variations within gene families were positively correlated with selected gene families in *T. gondii* strains across 14 chromosomes and reflected the significantly strain specific gene sets between varied isolates and the novel SNPs identified here compared to the previous finding done by Lorenzi *et al*., (2016).

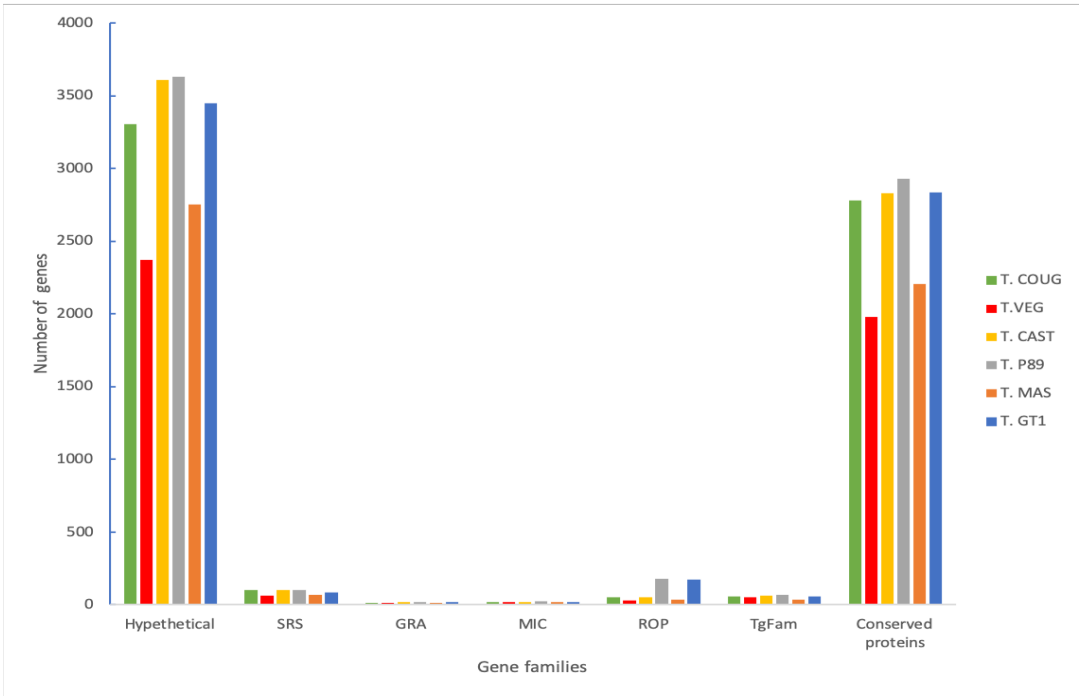



**Figure 5.12**: Frequency of relevant gene families with nonsynonymous SNPs in all the six strains of *T. gondii* as legend shown, includes surface (SRS), rhoptry (ROP), micronemes (MIC), dense granules (GRA), *Toxoplasma gondii* family (TgFAM; A, B, C, D and E). A large number of genes were annotated as hypothetical proteins. The lowest number were found in MICs and GRAs genes.

**Table 5.5**: The total variations numbers within expanded gene families in the *T. gondii* strains (SRSs, ROPs, GRAs, MICs, TgFAMs and KRUFs) that were enriched in SNPs based on the analysis of positively selected gene families in the *T. gondii* strains. The number between brackets referred to the novel SNPs identified in this study.

| Strain | SAG1-related-sequences (SRSs) | | Rhoptry kinase family (ROPs) | | Dense granules (GRAs) | | Micronemes (MICs) | | Toxoplasma gondii family (TgFAMs) | | Lysine-Arginine rich Unidentified Function family (KRUFs) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N. of SNPs | N. of genes | N. of SNPs | N. of genes | N. of SNPs | N. of genes | N. of SNPs | N. of genes | N. of SNPs | N. of genes | N. of SNPs | N. of genes |
| ***T. GT1*** | 1012 (676) | 83 | 433 (231) | 49 | 117 (89) | 16 | 136 (123) | 21 | 753 (690) | 58 | 150 (132) | 9 |
| ***T. MAS*** | 765 (432) | 64 | 301 (139) | 36 | 92 (67) | 13 | 68 (56) | 17 | 496 (400) | 37 | 81 (55) | 6 |
| ***T. P89*** | 1313 (876) | 102 | 531 (342) | 50 | 137 (101) | 18 | 151 (145) | 23 | 926 (890) | 67 | 129 (99) | 10 |
| ***T. CAST*** | 1238 (776) | 101 | 437 (289) | 52 | 142 (100) | 18 | 133 (122) | 21 | 749 (432) | 61 | 149 (85) | 9 |
| ***T. VEG*** | 779 (432) | 61 | 329 (291) | 25 | 123 (98) | 15 | 78 (66) | 17 | 807 (765) | 50 | 139 (120) | 7 |
| ***T. COUG*** | 1264 (870) | 99 | 279 (176) | 46 | 74 (53) | 13 | 80 (76) | 17 | 821 (768) | 57 | 99 (76) | 11 |

### 5.3.5 GO terms analysis

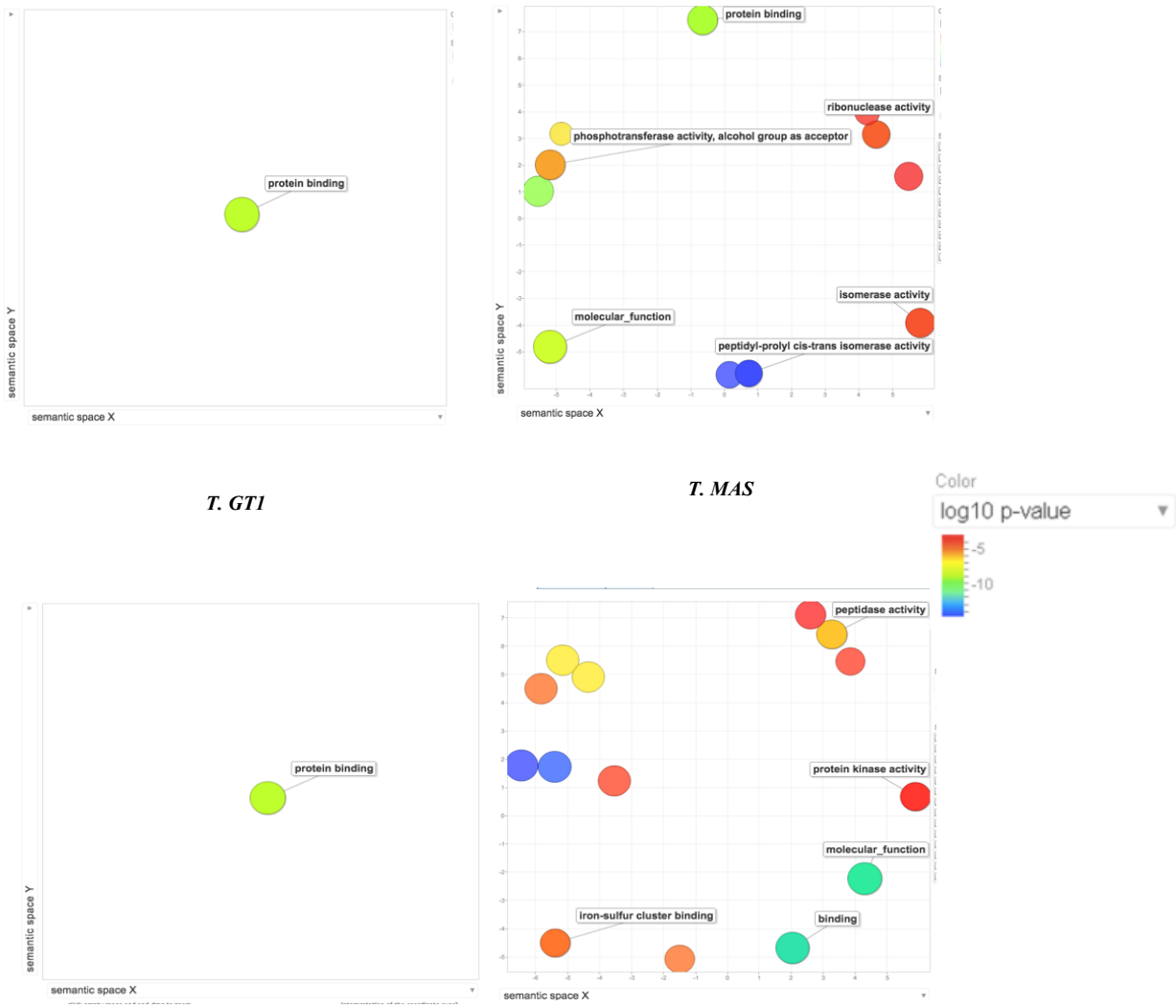
After performing functional annotations of the genes that harboured SNPs, we carried out GO term enrichment analysis and then visualised the analysis using REViGO. We found that different pathways were enriched within strains based on the number of SNPs grouped into varied impacts. This overrepresentation of some GO terms was important for increasing our understanding of the positive correlations between the specific pathways enriched and phenotypic changed encoded in the polymorphic genes identified.

#### 5.3.5.1 Pathways enriched in genes containing predicted modifier impact SNPs

In previous SNP analysis impact prediction, different impacts were found to be related to each strain of *T. gondii*. An impact comparison was made of all the mutations in all the strains that confirmed that the vast majority of the genes containing predicted SNPs modifier impact with greater than 89% in all strain. We expected that the greatest number of GO terms would be observed in genes encoding the modifying SNPs. In addition to this point, there was a dramatic increase in the number of GO terms that were similar between isolates. A total of 1,14,1 and 15 GO terms appeared in four strains of *T. gondii*; *T. GT1, T. MAS, T. CAST* and *T. VEG* respectively. No GO terms enriched were noticed in *T. P89* and *T. COUG* due to the lowest number of genes assigned. Most of the biological processes terms were related to the general metabolic processes that controlled different functions in the *T. gondii* isolates.

Significantly, the term protein binding is found over represented amongst genes containing predicted modifier impact SNPs in all strains. This reflects potential regulatory binding proteins belonging to the different group of genes within the SRSs, ROP, MICs, GRAs, TgFAMs and KRUFs families of genes and that play a significant function in pathogenesis and other biological traits shown in Figure 5.13. However, some additional GO terms were exclusively identified in specific isolates and confirmed the hypothesis of the selective pressure and genetic diversity among strains. As we can see from Figure 5.13, only one GO term (protein binding) was present in both *T. GT1* and *T. CAST* isolates, reflecting the high level of similarity between those

strains. In strain *T. MAS* and *T. VEG*, the pathways were more clustered with multiple GO terms such as kinase activity and peptide poly cis- trans isomerase activity that might be related to local variations and provide some general view of inter-protein variations that are likely to reflect divergent evolutionary paths based genetic and geographical diversity.




***T. GT1*** ***T. MAS***

***T. CAST*** ***T. VEG***

**Figure 5.13:** REVIGO's Scatterplot showing different cluster representatives of GO terms in genes containing modifying SNPs in four strains of *T. gondii*. No enrichment was found with in *T. P.89* and *T. COUG*. The clusters with larger blue and green circle have greater P- value assigned with more significant.

#### 5.3.5.2 Pathway enriched in genes containing predicted moderate SNPs

Examining the GO terms annotating genes with moderate SNPs, a large number were noticed in all the six strains 37, 33, 38, 36, 29 and 40 GO terms in *T. GT1, T. MAS, T. P89, T. CAST, T. VEG* and *T. COUG* strains, respectively (Figure 5.14). It has been found that the pathways of genes with this impact were clustered by their molecular functions in all the strains of *T. gondii*. GO terms were highly enriched in different aspects of function including protein binding, hydrolase activity, molecular transporter activity and transcription factors. Looking to the clustered GO terms between strains, the *T. COUG* isolate had the highest number of GO terms assigned to topics of protein binding, hydrolase activity and molecular transporter activity. A lower number of terms was observed in the *T. GT1, T. MAS, T.P89, T. CAST* and *T. VEG* isolates; this suggested that the genes containing SNPs with a predicted moderate impact are differed between strains.

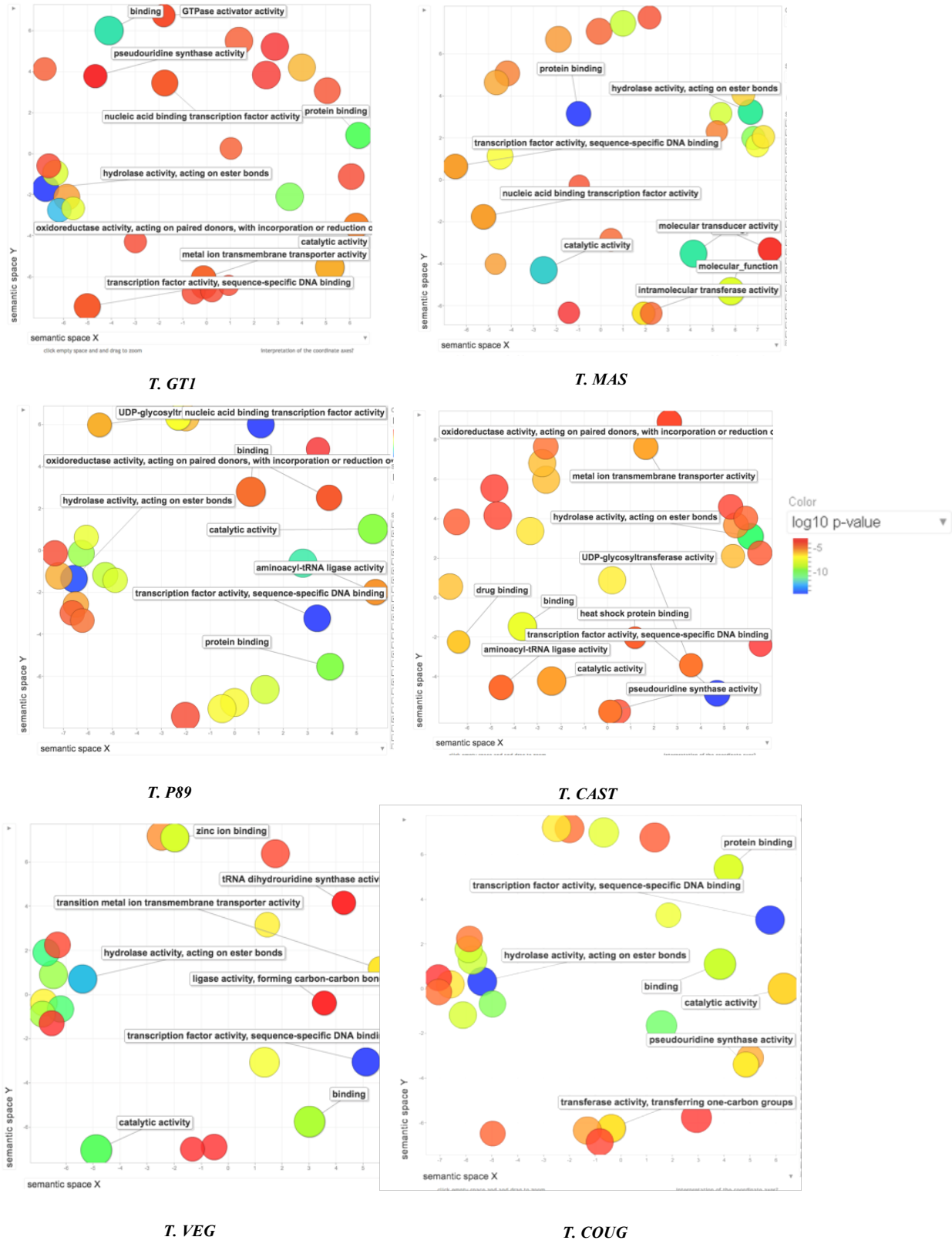
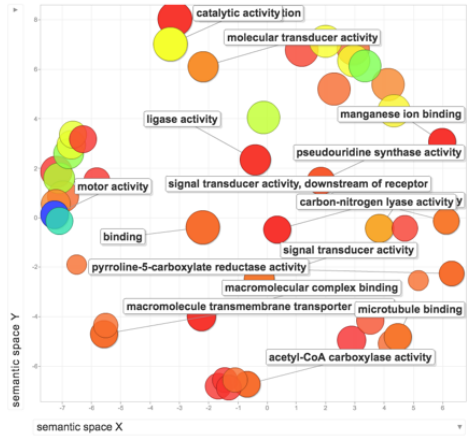



**Figure 5.14:** REVIGO's Scatterplot showing different cluster representation of molecular function ontologies for pathways of moderate impact SNP, summarised from different strains. Bubble colour indicates the p-value; size indicates the frequency of the GO term. The clusters with larger circle (blue and green) assigned with more significant.

#### 5.3.5.3 Pathway enriched in genes containing predicted high SNPs

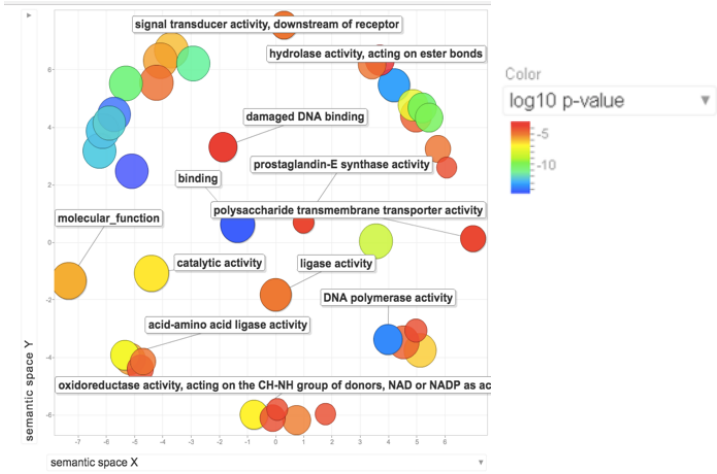
A total of 1496, 1146,1003,1567,629 and 1160 *T. gondii* genes were considered to be significant and had deleterious SNPs in the sequences of *T. GT1, T. MAS, T. P89, T. CAST, T. VEG* and *T. COUG* strains, respectively. Analysing the enriched terms, we found that there was a total of 63, 46, 72, 58, 19 and 57 terms in *T. GT1, T. MAS, T. P89, T. CAST, T. VEG* and *T. COUG* strains, respectively. Within the six strains, the highest enrichment proportion of GO terms was noticed in strain *T. P89* strain with 72 GO molecular terms followed by the *T. COUG* strain. However, a significant reduction in the number of GO molecular terms was highlighted in the *T. VEG* strain with only 19 GO terms. There were some molecular categories of the GO terms that showed similar patterns of occurrence shared between the isolates, which mainly involved specific functions including catalytic, hydrolase, transferase, motor activities and binding, which suggested that there were wide ranges of genes that play a similar role in the all six isolates (Figure 5.15). In addition to this, there were some GO terms that appeared in only some of the *T. gondii* strains, and these might reflect some strain-specific pathways related to clusters of genes with particular functions per strain. The proportions of GO terms in all the strains of *T. gondii* were plotted in Figure 5.16.



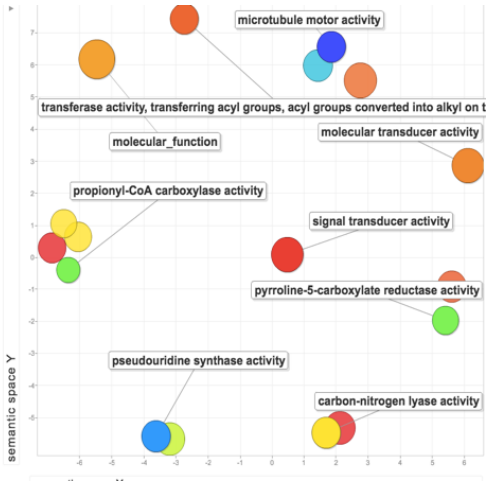

*T. GT1*


*T. MAS*

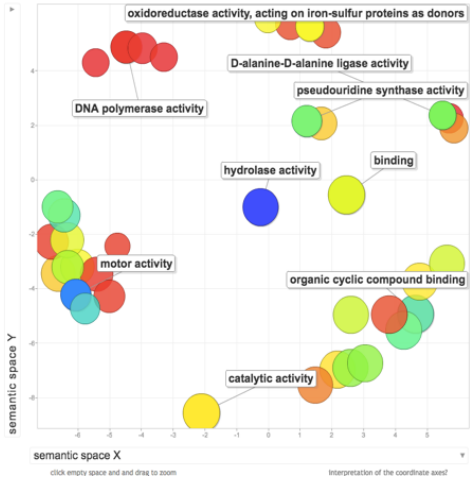

*T. P89*




*T. CAST*



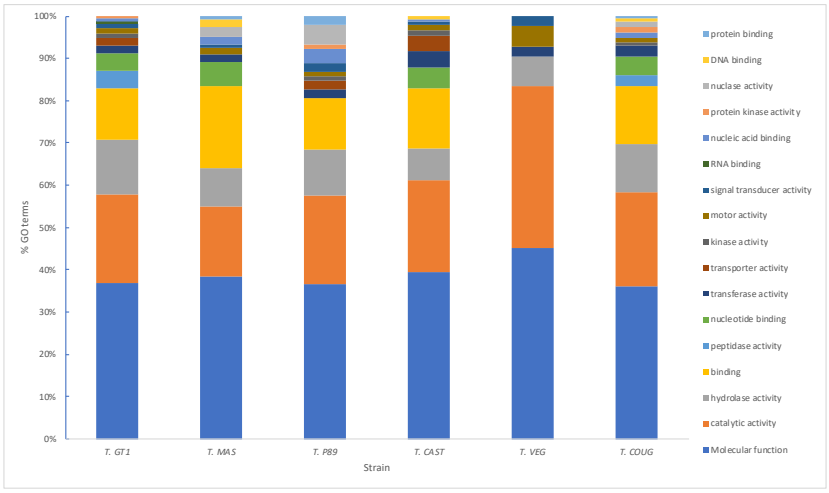

*T. VEG*



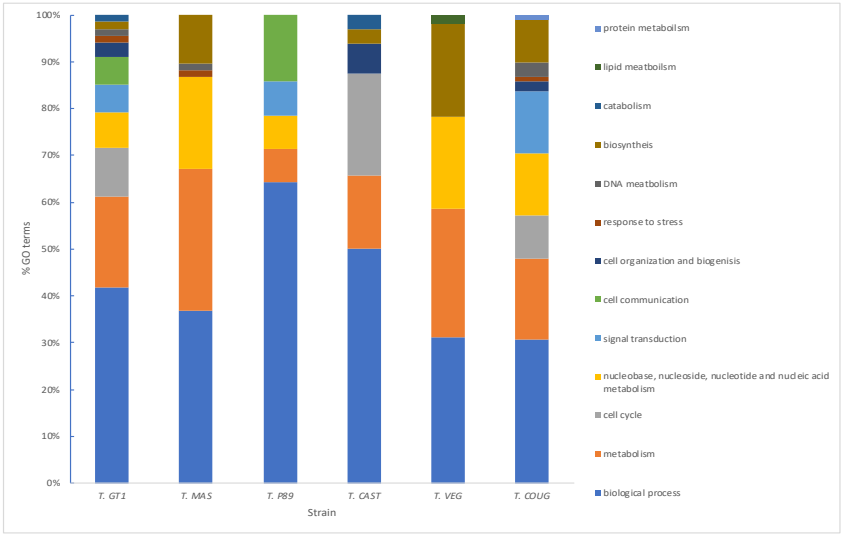

*T. COUG*

**Figure 5.15:** REVIGO's Scatterplot showing different cluster representatives of molecular functions ontologies of pathways of high impact SNPs were summarised from different strains Bubble colour indicates the p-value; size indicates the frequency of the GO term. The clusters with larger circle (blue and green) assigned with more significant.

A




B




C

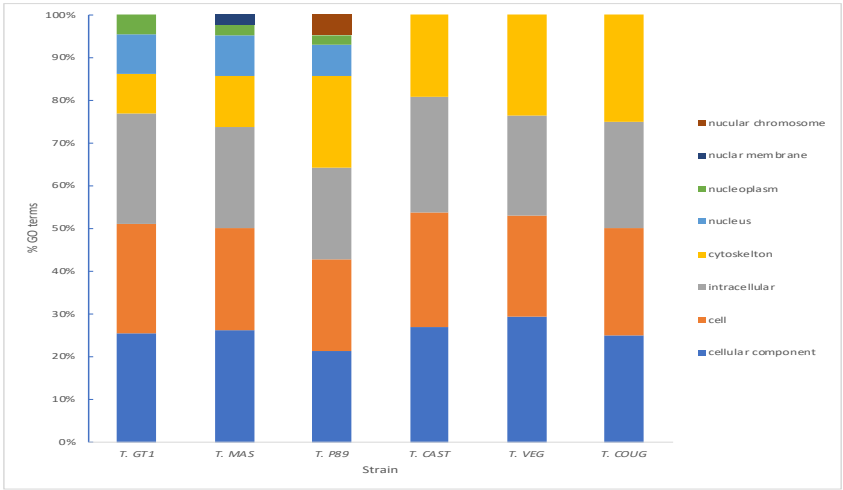



**Figure 5.16:** The proportions of GO terms that grouped into molecular functions (A), biological process (B) or cellular components (C) classification, showing the corresponding number of enriched GO terms such as having genes with high impact SNPs (X- axis indicates varied GO terms. The y- axis indicates the percentages of proportions of GO terms in specific term in all the strains of *T. gondii* as legend shown.

### 5.3.6 Investigating high impact SNPs unique to strains of *T. gondii*

We next went on to comprehensively catalogued the key genes bearing high impact SNPs as a consequence of SNPs in either stop or start codons (gain or loss) or splice sites. Those high impact SNPs were significantly lower in numbers, no more than 0.2% of the total SNPs observed. We noticed that *T.VEG* and *T. P89* were closely related strains. Analysis of strain-specific genes revealed that the *T. COUG* and *T. MAS* strains showed the highest proportions of unique genes that had deleterious effects (Table 5.6). A large number of frameshift effects were observed in *T. CAST* and *T. GT1* strains with 1522 and 1467 SNPs within 685 and 659 protein coding genes, respectively. Further dynamic changes in deleterious SNPs were also identified in predicted protein coding genes with SNPs that caused stop codons in the genes. The distinction of stop gained SNPs revealed that the *T. CAST* and *T. P89* had the highest proportions of SNPs in their genes with 683 and 670 SNPs within 539 and 573 genes respectively. The highest stop lost effects were noticed in the *T. P89* strain with 624 SNPs within 451 candidate genes. The details of the sub classification of the high impact SNPs per strain was plotted in Figure 5.17. There was a dramatic increase in the number of effects known as frameshift SNPs caused by insertion or deletion of bases in their sequences in that have a dangerous impact on protein effectiveness. When we are looking at stop gained frequencies between parasite isolates, we noted that the *T. CAST* acutely virulent strain had the highest number of stop gained with total of 683 SNPs located within 539 genes that led to early termination of gene expression due to different stop codons UGA, UAA and UAG generated. More than half of the total genes (53%) encoded predicted proteins of unknown function.

Of the six strains analysed, it has been found was lower occurrence of high impacts SNPs was in the *T. VEG* strain, with total number of 955 SNPs residing in 630 annotated genes. However, the strain *T. CAST* showed a dramatic increase in the number of high impact SNPs with 3019 SNPs distributed in 1567 annotated genes. To further examine the frequency of the high impact SNPs, we compared the frequencies in multiple locations between strains. It has been found that some annotated genes harboured more than 18 SNPs.

When expanding the comparison to include conserved genes, it was expected that the vast majority of the high impact SNPs were annotated as hypothetical proteins and

there was a varied number of high impact SNPs estimated per strain. See dataset of chapter five in appendices section named (C) include six *T. gondii* strains; C1(*T. GT1*), C2 (*T. MAS*), C3 (*T. P89*), C4 (*T. CAST*), C5 (*T. VEG*) and C6 (*T. COUG*) that have a high impacts of unique SNPs per strain.

**Table 5.6**: The number of genes with predicted high impact SNPs in each of the six strains of *T. gondii.*

| Strain | Number of predicted high impact SNPs | Number of genes |
|---|---|---|
| ***T. GT1*** | 1,496 | 2,853 |
| ***T. MAS*** | 1,146 | 2,238 |
| ***T. P89*** | 1,003 | 1,483 |
| ***T. CAST*** | 1,567 | 3,016 |
| ***T. VEG*** | 629 | 955 |
| ***T. COUG*** | 1,160 | 2,163 |

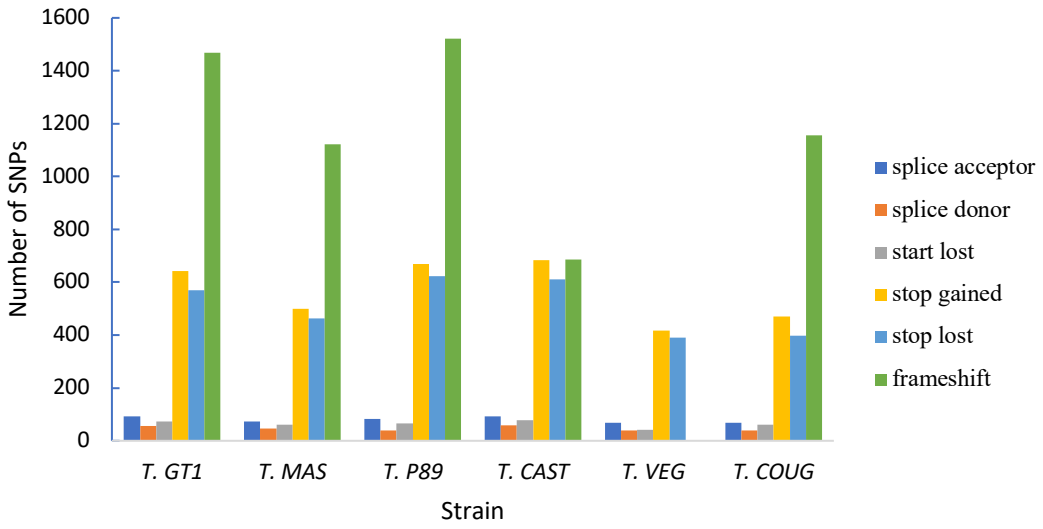



**Figure 5.17:** The contribution of each sub classification of high impact SNPs identified in all the six isolates of *T. gondii* as legend showed above. The maximum impact was observed as insertion or deletion causing a frameshift, then stop gained SNPs that causes stop codons, followed by stop lost that causes start codon to be mutated into a non-start codon.

### 5.3.6.1 High impact SNP identification in key biologically relevant gene families in distinct *T. gondii* strains

#### 5.3.6.1.1 Surface antigen gene family (SAG)

As previously discussed, members of known gene families were observed harbouring high numbers of SNPs, including; stop lost, stop gained, start lost and frameshift in the same gene. One of those group was the SAG gene family, known here as SRS proteins that are found in the membrane of the parasite and play a functional role in attachment and invasion. Such high impact SNPs in this specific group of proteins could resulted in loss of attachment to the host cell surface due to introducing stop codons that might cause truncation and loss of effectiveness of the protein product. This suggested that this specific family has some interesting features reflected in genetic variation within virulent and nonvirulent strains, which will be briefly explained below.

In the *T. GT1* strain, it has been found that there are 21 out of 83 SRS genes with high impact SNPs including six genes with a high number of deleterious SNP namely SRS48Q, SRS22I, SRS48E, SRS48K, SRS40B, and SRS26B. The highest number of SNP was noticed in SRS48Q totalling of 24 SNPs, followed by 11 truncating SNPs in the SRS22I protein. However, more than four high impact SNPs resided in the remaining four SRS genes. In *T. MAS* strain, 11 out of 64 genes were observed with a high frequency of high impact SNPs. We found the highest proportions of high impact SNPS in the SRS22I gene, totalling of 34 SNPs followed by 10 and 5 high impact SNPs residing in SRS15B and SRS16B, respectively, which are both located in chromosome IV. Further members of SAG-related genes were also noticed with 2 SNP for stop codons located in the SRS13, SRS30A, and SRS15C genes. In addition to those members, one high impact SNP was also observed in each of SRS40B, SRS22D, SRS30C, SRS20A and SRS12D distributed in different positions on different chromosomes. In *T. P89* strain, 10 out of 102 genes were observed with highly deleterious SNP causing a direct effect on the function of the proteins by generating new stop codons gain, lost, start lost, frameshift and also within the splice sites.

An examination of the members of SRS family that were consider paralogous genes demonstrated that there was one high impact SNP in SRS49C, SAG2D, SRS40B, SRS22D, SRS30A, SRS36A, SRS36C, SRS53D, SRS12D, SRS10, SRS51 and SRS3. For the detailed functionality of proteins carrying the 10 deleterious SNPs, identification of the functional consequences of truncating SNPs in the genes was performed. Given the observation that those SNPs can be associated with predicted pathogenic genes, we considered 4 stop gains, 2 stop lost, 2 start lost and finally 2 splice acceptor SNPs affecting ten of the major surface antigen. However, no frameshift SNPs was noticed in this group of the coding protein genes. Importantly, stop gained SNPs were found in a subset of genes (SRS36A, SRS36C, SRS10 and SRS53D genes) that were thought to be involved in signalling pathways, host parasite interaction, attachment and localized in the membrane of the parasite. Analysis of further GPI-anchored proteins also revealed that these could be responsible for some of the inter-strain variation, more specifically in the interaction between host and parasite. Two members of this family had one predicted stop loss SNP in SRS40B and SRS22D respectively. Additionally, four deleterious SNPs occurred in SRS49C and SRS12D in addition to two start lost and two splice acceptor site SNPs in the SRS10 and SRS51 genes respectively.

A total of 81 nonsense SNPs were identified in 21 out of 101 genes that had distinct categories of significant dangerous effects in the *T. CAST* strain. A similar pattern of selection of high impact SNPs was noticed in the *T. CAST* and *T. GT1* strains. Within this group of genes there was a dramatic increase in the number of deleterious mutations in the same two SAG-related sequence genes SRS48Q and SRS12D totalling 24 and 8 frameshift SNPs respectively. In SRS12D, there was one further stop gained SNP. In addition, one or SNPs were detected within the remaining 18 SRS members including SRS16B, SRS22G, SRS22I, SRS48E, SRS15C, SRS52F, SRS48K, SRS49B, SRS13, SRS55N, SRS15B, SRS49C, SRS18, SRS40B, SRS22D, SRS31B, SRS30A, SRS51 and SRS53D. It was shown that most of the SNPs detected were functionally classified into frameshift mutations leading to nonsynonymous amino acid changes and phenotyping diversity in the *T. CAST* strain.

As might be expected, the proportion of high impact SNPs noticed in the *T. VEG* strain was significantly lower than the values observed in other strains of *T. gondii* see Table 5.4. We detected four SNPs in four members out of 61 SRS family genes that contained SNPs according to table 5.7; SRS49C, SRS40B, SRS36A and SRS36C with one high SNP grouped into different functional predicted effects types including start lost, stop lost and stop gained SNPs. Significantly, *T. COUG* had the highest number of SRSs genes affected with 30 out of 99 SRS genes that contained predicted SNPs. These were annotated as SRS48Q, SRS48K, SRS49A, SRS49B, SRS18, SRS40D, SRS40B, SRS45, SRS22C, SRS22D, SRS22G, SRS22I, SRS59K, SRS26E, SRS26B, SRS26A, SRS31B, SRS30D, SRS30A, SRS47A, SRS19B, SRS19F, SRS51, SRS55N, SRS16E, SRS16B, SRS15C, SRS15B, SRS12D and SRS12B. Of particular of interest, the highest level of polymorphism was identified in five the SAG1-related family of surface antigens including SRS22I with 26 SNPs and both SRS22G and SRS48K with 17 SNPs each. Nine SNPs were found within SRS16B and a further 8 deleterious mutations were noticed in the SRS22D gene.

A comparison of the six *T. gondii* isolates revealed that there were differences in the number of high impact SNPs identified in the SRS gene families as summarised in Table 5.7. It is evident that the *T. P89* and *T. COUG* strains had the greatest number of mutations totalling 1313 and 1264 SNPs distributed in 102 and 99 SRS genes respectively. Overall, our SNP prediction analysis demonstrated that the number of SNPs that were positively correlated with the number of genes affected by deleterious SNPs. According to our study, a total of 125 out of 1264 SNPs were considered high impact in 30 members of the SRS family in the *T. COUG* strain followed by the *T. CAST* and *T. GT1* strains with 81 and 79 SNPs located in 21 SRS genes respectively.

Interestingly, the highest proportion of the high impact SNPs was noticed in the SRS22I gene with 34 and 26 SNPs in *T. MAS* and *T. COUG* isolates. A further of the genetic diversity in this particular gene family was made between strains, which demonstrated that there were 24 high impact SNPs out of 79 and 81 SNPs were found in SRS48Q in the *T. GT1* and *T. CAST* strains respectively. Significantly, there were dramatically fewer high impact SNPs in the *T. VEG* strain. The SRS genes with predicted high impact SNPs are summarized in Figure 5.18.

#### 5.3.7.1.2 Rhoptry kinase proteins ROPs

Analysis of further gene families for members unique to each representative isolate revealed that there were dramatic difference in the functional characterization of the rhoptry kinase proteins and rhoptry neck proteins in terms of their high impact SNPs. Genes were identified that showed pathogenic factors carrying different patterns of SNPs that were considered to be deleterious namely ROP1, ROP2, ROP5, ROP8, ROP16, ROP18, ROP35, RON2, RON5 and RON8 (Figure 5.19). Interestingly, the frequencies of different deleterious effect categories varied per strain, although, it has been found that there were some conserved SNPs in all strains. Amongst the ROPs proteins, we also noticed further members of rhoptry neck proteins named (RONs) that were involved in the moving junction complex during invasion infection processes. This complex consists of RON2, RON4, RON5 and RON8 that interact with microneme protein AMA1 (Takemae *et al.*, 2013).

From the Table 5.7, a total of 10 rhoptry candidate genes were identified in the *T. GT1* strain with high impact SNPs. Seven of these genes contained 15 high impact SNPs in the ROPs; ROP15, ROP8, ROP16, ROP5, ROP1 and rhoptry kinase family protein (ROPK), ROP19A and ROP19B. ROP8 had the highest number of the SNPs totalling 6 high impact SNPs located in chromosome X, 5 of which were considered as frameshift SNPs and one stop lost. Two frameshift SNPs resided in ROP19B, ROP16 and ROP1. Additionally, one frameshift SNP was found in each of ROP15, ROP5 and ROP19A.

More importantly, ROP5 and ROP16 had frameshift SNP, which would result in a non-functional protein product and might be associated with the reduction of virulence, depending on the strain of the *T. gondii*. In terms of pathogenicity, ROP5 and ROP16 were both highly active in type I strains, however ROP5 was minimally active in type II & III. Furthermore, ROP16 was highly active in type III. Functionally, those two proteins contribute to regulate host signalling in the STA3/6 activation pathway and in manipulation of the host immune response by interacting with other secreted ROP kinase proteins and hence a nonsynonymous changes may have an impact on the virulence. ROP19A was noticed with one stop gained SNPs which plays a key role in the parasitophorous vacuolar membrane (PVM).

Five high impact SNPs were also observed in the RON family, more specifically, in RON2, RON5 and RON8. The highest number of the high impact SNPs in RON2 with 2 stop gained SNPs followed by one stop gained SNP noticed in the RON8 gene with one additional start lost SNP in the RON5 gene.

In *T. MAS* strain, nine high impact SNPs were located in five rhoptry kinase proteins and one rhoptry neck protein (ROP15, ROP16, ROP19A, ROP19B, ROP35 and RON8 proteins). The greatest number of SNPs that were assumed to have a high impact on the parasite's pathogenic genes was noticed in ROP15, ROP16 and RON8 with two stop gained and loss SNP in each of the ROP15 and RON8 genes, respectively, as well as, two frameshift SNPs in ROP16. The remaining deleterious SNPs were in ROP19A with stop gained, frameshift in ROP19B and start lost in ROP35 proteins. Remarkably, and similar to *T. GT1*, the two frameshifts SNPs in the ROP16 protein cause a non-functional gene. Amongst the secreted rhoptry proteins, ROP15 and ROP35 were further identified with one stop gained and one start lost SNP respectively, which might have cellular functions in cell cycle and oocysts development (Wang *et al.*, 2017).

As we mentioned earlier, several members of the RON family (RON2, 4, 8) are conserved between the two strains. This finding suggests that these high impact SNPs were primarily related to conserved invasion, moving junction and host immunity manipulation processes among *T. gondii* strains. In addition to this, more members of ROPs and RONs were investigated in the *T. gondii* strain P89. Four significant high impact SNPs were observed in a subset of the ROP genes. Genes observed with those effects were grouped into two stop gained mutations in each of ROP15 and ROP19A. In addition, more SNPs identified to be deleterious with one stop lost effect localized to ROP8 and one start loss in ROP39. The same pattern of the high impact SNPs in RONs members was found in RON2, RON8 and RON5 with one stop gained and also a loss SNP in RON8 and one stop loss SNP in each of RON2 and RON5 respectively.

A total of 13 out of 52 ROP genes were found to be harbouring 25 high impact SNPs in their sequences. Interestingly, it has been found that there was a similar within the *T. GT1* and *T. CAST* strains, suggesting that the two strains have some conserved sequences shared between them, despite variation in the types of the effects of the deleterious SNPs. The prediction of the SNPs effects on the ROP and RON proteins

revealed that most of the pathogenicity genes have high impact SNPs, which we have previously mentioned in *T. GT1* strain, including ROP15, ROP8, ROP2A, ROP19A, ROP19B, ROP16, ROP5, ROP1, RON2, RON8 and RON5 genes. In the *T. CAST* strain, 5 out of 25 high impact SNPs were noticed in the ROPK family including in ROP8 with a frameshift effect. We further classified the rest of the truncation SNPs into different effects including the tandemly triplicated ROP2 gene known here as ROP2A with two frameshift SNPs. The emergence of the nonsense mutations in ROP2A might be not associated with host mitochondria as highlighted previously (Kemp, Yamamoto and Soldati-Favre, 2013). In addition to ROP2A, ROP35 and RON9 were identified with one start loss SNP and two frameshifts in ROP35 and RON9. Remarkably, most of the high impact SNPs in this strain represented predicted frameshifts; this suggested that these might impact the function of this rhoptry gene, which is a vaccine candidate due to the highly protective mechanism, importantly, also in the highly virulent strains belonging to the type I as confirmed in *in vitro* experiments reported by (Chen *et al.*, 2014).

Fewer high impact SNPs have been observed with stop gain and start lost in ROP15 and ROP35 proteins respectively with an average of one functional effect each in the *T. VEG* strain. Interestingly, it seems that there was no frameshift SNP in this specific strain. In the *T. COUG* lineage, a total of 8 out of 46 encoded genes contained 8 SNPs in genes of the ROP family. The comparison between ROP and RON genes indicated that a subset of rhoptry proteins have a different number of SNPs, including ROP8, ROP7, ROP4, ROP35, ROP5, ROP1 and in two rhoptry neck proteins, RON9 and RON5. It has been noticed that the most of the functional effects of the deleterious SNPs here were enriched with frameshift effects. Only one stop and start loss SNP was observed in the sequence of RON5 and ROP35, respectively. It was hence clear that the repertoires of the ROPs gene that had high impact SNPs differed greatly depending on the lineages of the *T. gondii* pathogen. From our analysis, we propose that such high impact SNPs might be associated with variations between global *T. gondii* strains.

As we confirmed earlier in section 5.3.3.4, most of the investigated ROP genes that contained SNPs where compared to the refence genome of *T. gondii* strain ME49 belonged to the ROP5, ROP16, ROP17, ROP18, ROP39, RON3 and RON8 genes. The highest proportions of high impact SNPs were noticed in the *T. CAST* strain with a total of 13 ROP genes affected followed that by *T. GT1* and *T. COUG* strains with a total of 9 and 8 ROP genes, respectively.

Similar strain specific numbers of ROP proteins were observed in the two *T. gondii* strains *T. MAS* and *T. P89* totalling 6 ROPs encoding proteins. However, only a small number of ROPs genes was noticed in the *T. VEG* strain. This provided confirmation that strain *T. CAST* was the most diverse strain when compared to the others. The most striking result to emerge from the data was the highly significant enrichment of high impact SNPs with frameshifts effects in ROP8 protein in three isolates *T. COUG*,*T. GT1* and *T. CAST* strains with a total of 7, 6 and 5 SNPs each. The ROP genes with predicted high impact SNPs are summarized in Figure 5.19.

**5.3.7.1.3 Dense granules (GRA) genes:**

Generally, we found that GRA genes had the lowest proportions of the high impact SNPs compared with the other five distinct gene families (see Table 5.7). One high impact SNP was observed in the GRA3 gene in three strains of *T. gondii*; *T. GT1, T. P89* and *T. VEG* respectively. In addition to GRA3, one additional member of this family had a high impact mutation located in GRA10 in the *T. MAS* strain. Three dense granules proteins, GRA3, GRA10 and GRA15, were observed in *T. CAST* strain with three high impact SNPs each. A large number of GRA family members was identified in *T. COUG* with six mutations residing in four members of the GRA family; GRA1, GRA4, GRA15 and GRA5. Significantly, GRA3 was the only variable protein that was located in chromosome X in five strains of *T. gondii* with the exception of the *T. COUG* strain. Importantly, detection of the high impact mutations in the GRA1, GRA3 and GRA15 proteins contributed to modification of the PVM after cell invasion and also *T. gondii* survival.

#### 5.3.7.1.4 Micronemes genes (MICs)

In addition to the three gene families described above, the six strains of *T. gondi* revealed a set of apical organellar genes that are involved in attachment and actomyosin machinery known as micronemes genes (MICs). Of these MIC genes, five microneme proteins were identified in most of the *T. gondii* strains affected with high impact SNPs including the duplicated gene MIC17 with two copies (MIC17A and MIC17B) and also MIC11, MIC12, and MIC15 (Table 5.7). The microneme proteins contain a thrombospondin repeat (TSR) and are likely to be involved in adhesion and gliding motility by interacting with members of the rhoptry proteins during moving junction processes between the host cell and the parasites (Sheiner *et al.*, 2011).

Our analysis highlighted high impact SNPs in microneme protein MIC11 in all strains. The function of this gene was in controlling Toxoplasmosis disease as a vaccine target for this pathogen. The highest number of SNPs was observed in the *T. GT1* strain, totalling six SNPs. However, a low number of MIC genes was noticed in *T. VEG* strain with only one high impact SNP. Most of the functional effects were categorised as frameshift and stop gained effects that are predicted to be deleterious.

#### 5.3.7.1.5 Toxoplasma gene family (TgFAMs)

A comparison of the TgFAMs, divided into subgroups TgFAMA-E, highlighted the second largest proportion of the high impact SNPs from diverse isolates. Members varied significantly from strain to strain and also in the frequency of high impact SNPs in the candidate genes (see Table 5.7). Interestingly, the vast majority of TgFAMs genes belonged to the group TgFAMA in all the six strains. Two strains showed high number of SNPs with 68 and 56 SNPs in 20 and 11 TgFAM genes of the *T. P89* and *T. MAS* isolates, respectively. However, the highest number of candidate genes was noticed in *T. GT1* with a total of 22 out of 58 TgFAMs genes which may indicate a correlation between the frequencies of SNPs and the type of strain. Three out of six isolates showed a notable paucity of high impact mutations in some members of the TgFAME family with 30, 29 and 14 SNPs in *T. MAS*, *T. CAST* and *T. COUG* strains respectively in chromosome IV. This might be explained as some specific diversification and gene expansion clustered as TgFAME genes in particular strains.

#### 5.3.7.1.6 Lysine- Arginine rich Unidentified Function family (KRUFs)

Significantly, all the members belonged to the Lysine- Arginine rich Unidentified Function family known here as the KRUF genes family. It had the lowest frequencies of high impact SNPs in all six strains of *T. gondii.* A total of 27 SNPs were observed in *T. GT1* and *T. CAST* strains in 10 and 9 genes, respectively. Our analysis indicated an expansion of KRUF genes in haplogroups 1 and 7 (represented by T. GT1 and T. MAS). Functionally, it has been reported that KRUF family proteins are considered an example of merozoite unique gene families whose members were highly stage regulated (Hehl *et al.*, 2015). The frequency of gene families among genes with high impact SNPs between *T. gondii* strains was plotted in Figure 5.20 below.

**Table 5.7**: The total number of expansion gene families in *T. gondii* that contained high impact SNPs in different gene families that were uniquely enriched and expanded based on the analysis of positively selected gene families in different *T. gondii* strains.

| Strain | SAG1-realted-sequences (SRSs) | | Rhoptry kinase family (ROPs) | | Dense granules (GRAs) | | Micronemes (MICs) | | Toxoplasma gondii family (TgFAMs) | | Lysine-Arginine rich Unidentified Function family (KRUFs) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No. of SNPs | No. of genes | No. of SNPs | No. of genes | No. of SNPs | No. of genes | No. of SNPs | No. of genes | No. of SNPs | No. of genes | No. of SNPs | No. of genes |
| ***T. GT1*** | **79** | **21** | **19** | **10** | **1** | **1** | **10** | **3** | **41** | **22** | **27** | **9** |
| ***T. MAS*** | **60** | **11** | **9** | **6** | **2** | **2** | **2** | **1** | **56** | **11** | **14** | **3** |
| ***T. P89*** | **10** | **10** | **8** | **7** | **1** | **1** | **6** | **3** | **10** | **9** | **1** | **1** |
| ***T. CAST*** | **81** | **21** | **25** | **13** | **3** | **3** | **6** | **3** | **68** | **20** | **27** | **10** |
| ***T. VEG*** | **4** | **4** | **2** | **2** | **1** | **1** | **1** | **1** | **10** | **8** | **1** | **1** |
| ***T. COUG*** | **125** | **30** | **26** | **8** | 6 | 4 | **5** | **3** | **44** | **14** | **19** | **5** |

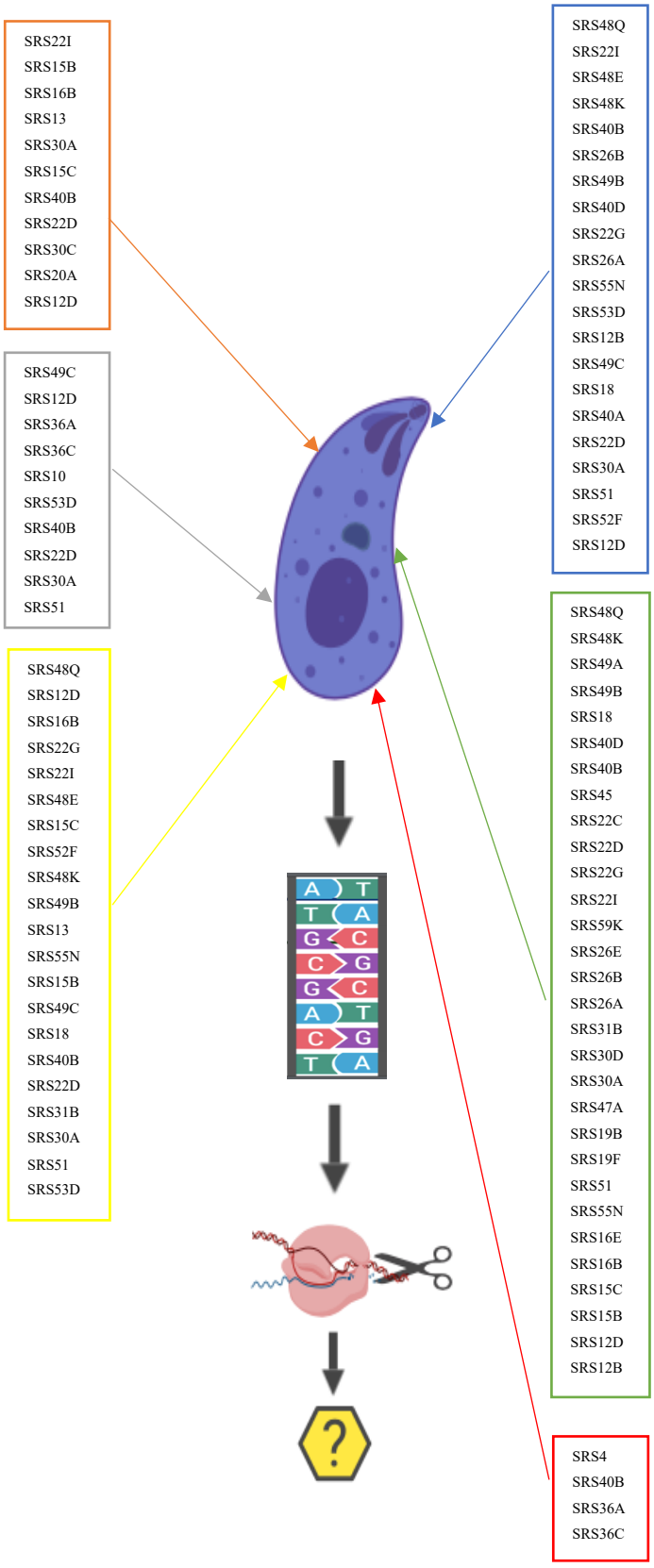



**Figure 5.18:** The variation in the SRS genes that had predicted high impact SNPs in the six linages of *T. gondii* resulting in non-functional proteins. Each strain is depicted in a different colour box, listing the subset of ROP and RON genes identified. Blue box (*T. GT1*), orange (*T. MAS*), grey (*T. P89*), yellow (*T. CAST*), red (*T. VEG*) and green (*T. COUG*). The gene names ranked from the highest number of SNPs to the lowest.

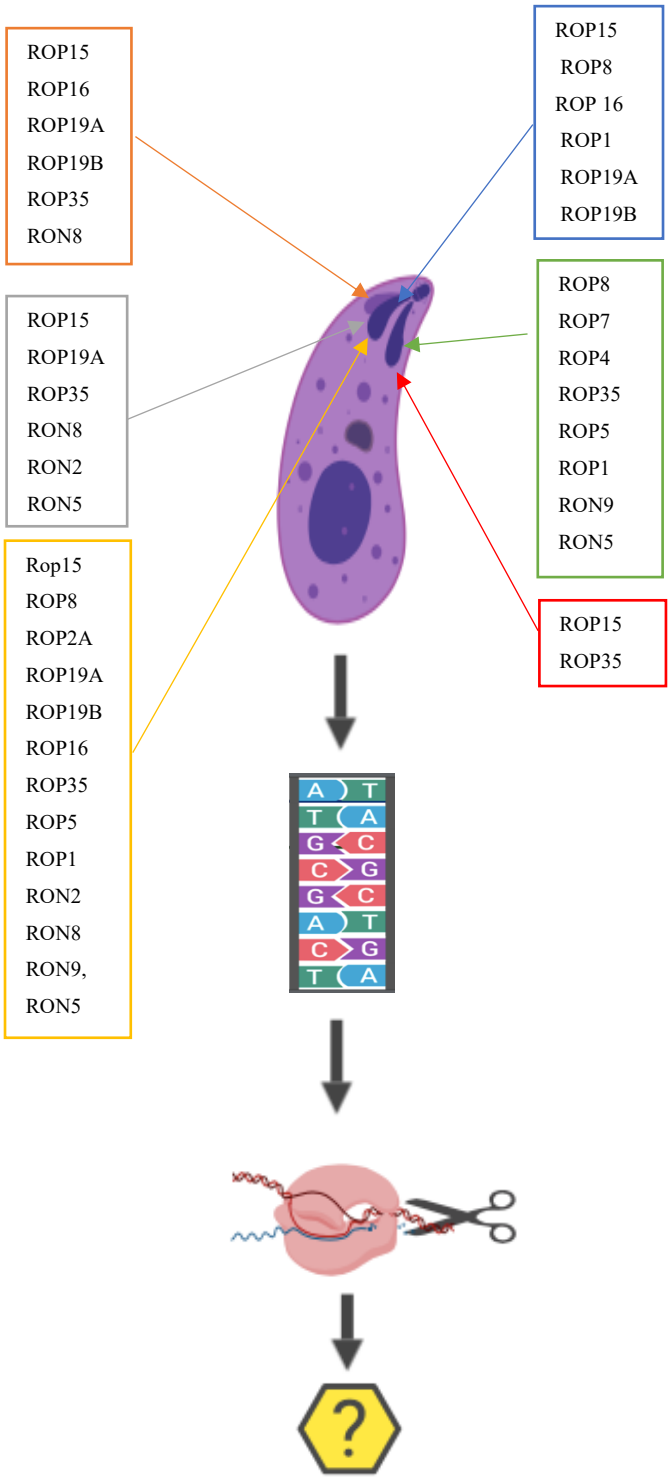



**Figure 5.19:** The variation in the rhoptry kinase genes (ROPs & RONs) encoding high impact SNPs resulting in non-functional proteins. Blue box (*T. GT1*), orange (*T. MAS*), grey (*T. P89*), yellow (*T. CAST*), red (*T. VEG*) and green (*T. COUG*). The gene names ranked from the highest number of SNPs to the lowest.

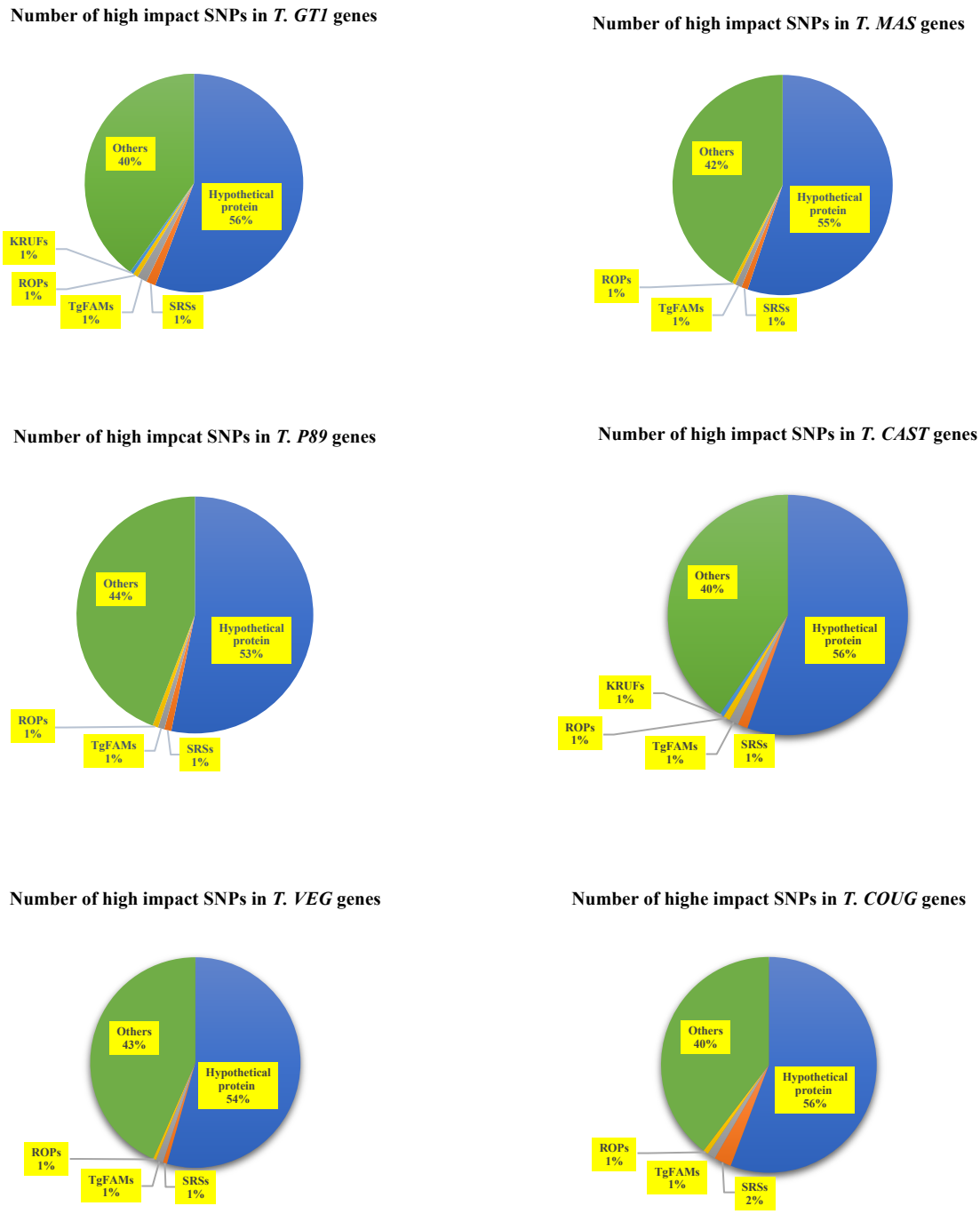



**Figure 5.20:** The frequency of gene families among genes with high impact SNPs between *T. gondii* strains including varied gene families; Rhoptry (ROPs), *Toxoplasma gondii* family A, B, C, D and E and Surface antigen proteins (SRSs).

### 5.3.7 Copy number variation (CNV) influences genomic diversity amongst *T. gondii* strains

The number of duplications and deletions at genomic loci represented an additional source of variation between the different strains. In order to investigate this, the CNVentor software was used. We examined the whole genomic sequences of the six isolates to identified CNVs of one kilobase (kb) or larger. As we described in Chapter 2, the sensitivity of the CNV analysis was related to the bin size chosen in this study namely 1000 bp. It has been found that increasing the bin size will allow prediction of fewer deletion and duplication CNVs. We identified duplicated and deleted sequences that were mainly distributed in telomeric and subtelomeric regions, as we expected, across the 14 chromosomes in all strains.

Overall, we discovered that the vast majority of CNVs are deletions compared to the reference genome sequence. For the isolates *T. GT1, T. MAS, T. P89, T. CAST, T. VEG* and *T. COUG*, the deletions number and the length were 4546, 2006, 5566, 2542, 5273, 3421 and 360, 140, 960, 250,480 and 190 kb respectively. The number of deletion events were significantly higher than duplications, more significantly in the *T. P89* strain with a total of 9600000, most of which were on five chromosomes ( VIIb, XII, XII, V and VI). More specifically, the highest number of CNV deletions was found in chromosome XII with a total of 153 kb.

We noticed that CNV duplications were mostly located in telomeric regions of most chromosomes in all strains of *T. gondii*. The largest proportions of CNV duplication was found in the *T. GT1* strain and were mainly located in telomeric regions of the IX chromosome with a total length of 185kb; (see Figure 5.21). More specifically, we found that there was a difference in the number of the CNVs between the six strains of *T. gondii* as shown in Figure 5.22. We further examined the gene contents of the segmental duplications that might contain some virulence related genes that evolved rapidly and have functional consequences on the pathogenicity differences between strains. Somewhat expected, the subset of protein-encoding genes that contain paralogous genes from multiple gene families included the SRSs, ROPs, MICs, GRAs, TgFAMs and KRUFs.

In *T. GT1* strain, the largest number of CNV was noticed in *Toxoplasma gondii* family proteins known as TgFAMA, which include three members clustered in the telomeres of the chromosomes XII. A total of 7 TgFAM genes were found duplicated on chromosome XII belonging to the TgFAM A and TgFAM B families annotated as TGME49_278080, TGME49_278090 and TGME49_278100 (Figure 5.23).

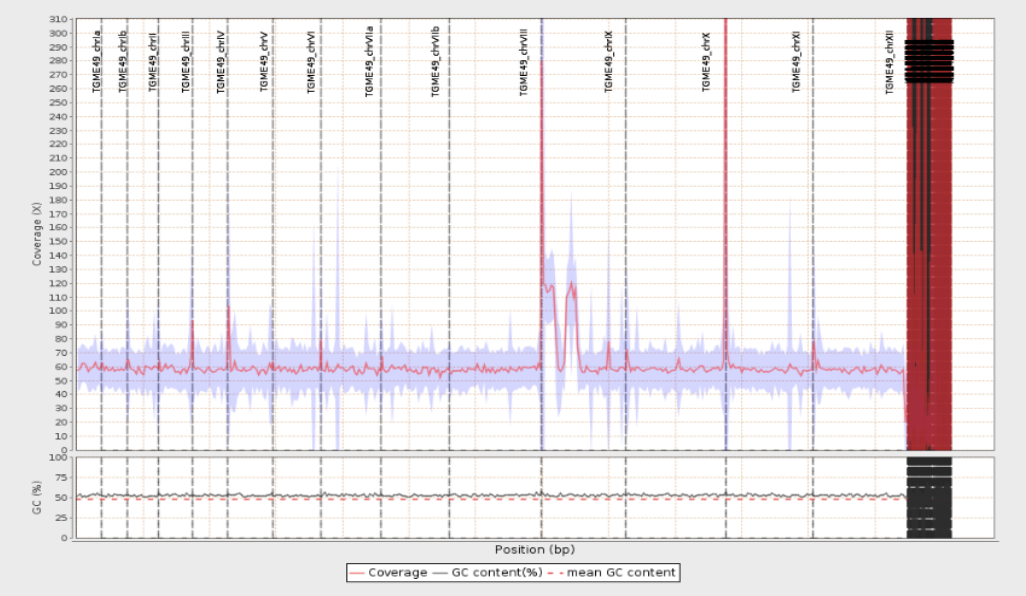
We also noticed that the second largest group of amplified genes were 12 SRSs genes making up only 11% out of the total 111 SRSs genes in the reference genome ME49, encoded on chromosome IX. Examination of the pattern of the CNV here revealed that 6 out of 12 SRS were tandemly duplicated genes known as SRS38A, SRS38B, SRS38C, SRS36C, SRS36D and SRS36E. (Figure 5.24 A & B). In the same chromosome, we also noticed that there were additional members that belonged to other gene families, specifically, three members of the KRUFs with CNV annotated as TGME49_210600, two of which were tandemly duplicated (TGME49_292390-TGME49_292375). MIC12 was also identified. Additionally, a gene annotated as 'lipoprotein, putative' was observed with a CNV and we hypothesize that this might play an important role in cholesterol pathways that have been observed as important for the infection process (Coppens, Sinai and Joiner, 2000). Other family genes were found including 6 ROP members annotated as ROP5, ROP7, ROP8, ROP2A, ROP26 and ROP19B, Dense granule family (GRA11) and also three micronemes (MICS) include MIC12, MIC16.

In *T. MAS* strain, the XII chromosome had fewer genes with CNV only 35 genes and most of them were of unknown function (hypothetical proteins). One of those had more than 20 copies in the hypothetical protein (TGME49_300790). In V chromosome, 9 out of 28 genes with CNVs, annotated as encoding hypothetical proteins. Furthermore, the most abundant gene families identified in both chromosomes in both strains were SRS, ROPs and TgFAMs (15 TgFAMA, 2 SRS20 A and C and 2 MIC8 and 9).

Interestingly, one member of the ROP family appeared in both strains in the same chromosome with a high number of copies, ~7 copies annotated as TGME49_308090 named as ROP5 that was tandemly duplicated in *T. MAS*, however only 3 copies were noticed in *T. COUG* strain.

Another gene family which was identified was TgFAMA with more than 12 members arranged as tandemly duplicated genes at the end of chromosome XII. In chromosome V, the SAG-related sequences SRS20C, SRS20A were noticed in both strains. A second large group of CNV duplications was noticed in *T. P89* and *T. CAST* with an identical number of CNV duplications of length 54000bp in chromosome XII, confirming that there were multi-copy coding genes that we predicted to cause the CNV duplication in this specific locus. In agreement with previous findings in *T. MAS* and *T. COUG*, this regions revealed that ROP5 was the most likely candidate between the four strains compared to the reference genome with different number of copies observed (Figure 5.25) Lastly, we found no significant evidence of duplication in the *T. VEG* strain; this may indicated that the low coverage of reads obtained for this strain might have impacted on our ability to detect CNVs with currently available software tools.

A




B

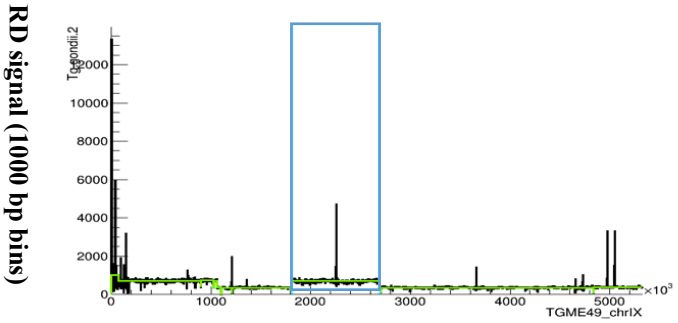



**Figure 5.21**: **A)** The coverage histogram across the 14 chromosomes of reference genome *T. gondii* strain ME49. In *T. GT1* strain there were two peaks in chromosome IX. This histogram was generated by using Qualimap 2 tool (http://qualimap.bioinfo.cipf.es). **B)** An example of the largest duplicated region with CNV length= 879000 bp in *T. GT1* strain marked in blue rectangle including different gene families includes SRSs, TgFAMA and KRUFs members. The y axis is the location per bp and the X axis was the read depth (RD) for predicted CNVs and the bin size was 1000 bp. The black histogram is indicated to the Read Depth (RD) signal for the fragment of chromosome IX. Green line is partitioning by CNVnator programme (http://sv.gersteinlab.org/cnvnator/)

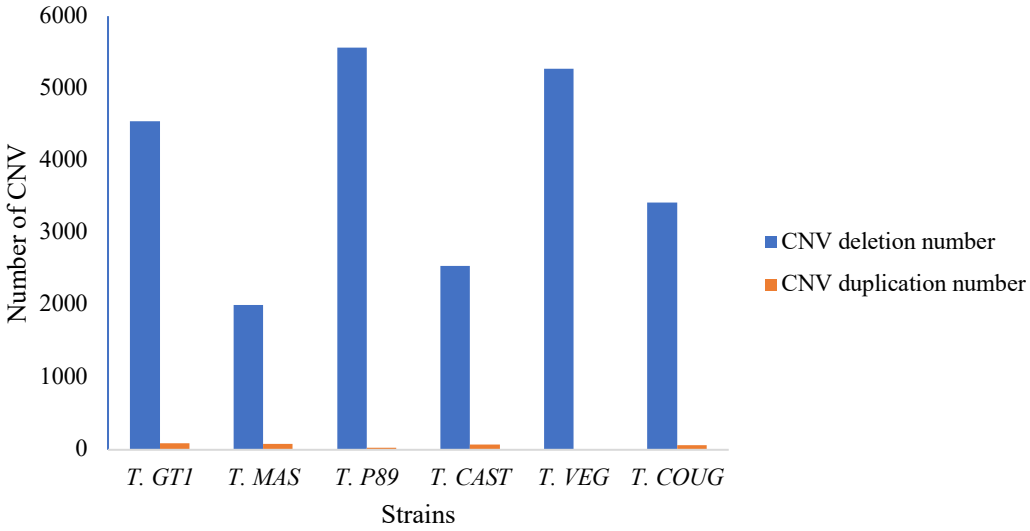



**Figure 5.22:** A comparison of CNV duplication and deletions across the genomes of the six strains of *T. gondii*, using CNVentor tool. The minimum CNV deletion was observed in *T. CAST* and *T. MAS* isolates, the maximum was in *T. P89* strain. A high number of duplications was noticed in *T. GT1* strain due to the high coverage obtained. No CNV duplications were seen in *T. VEG* strain.

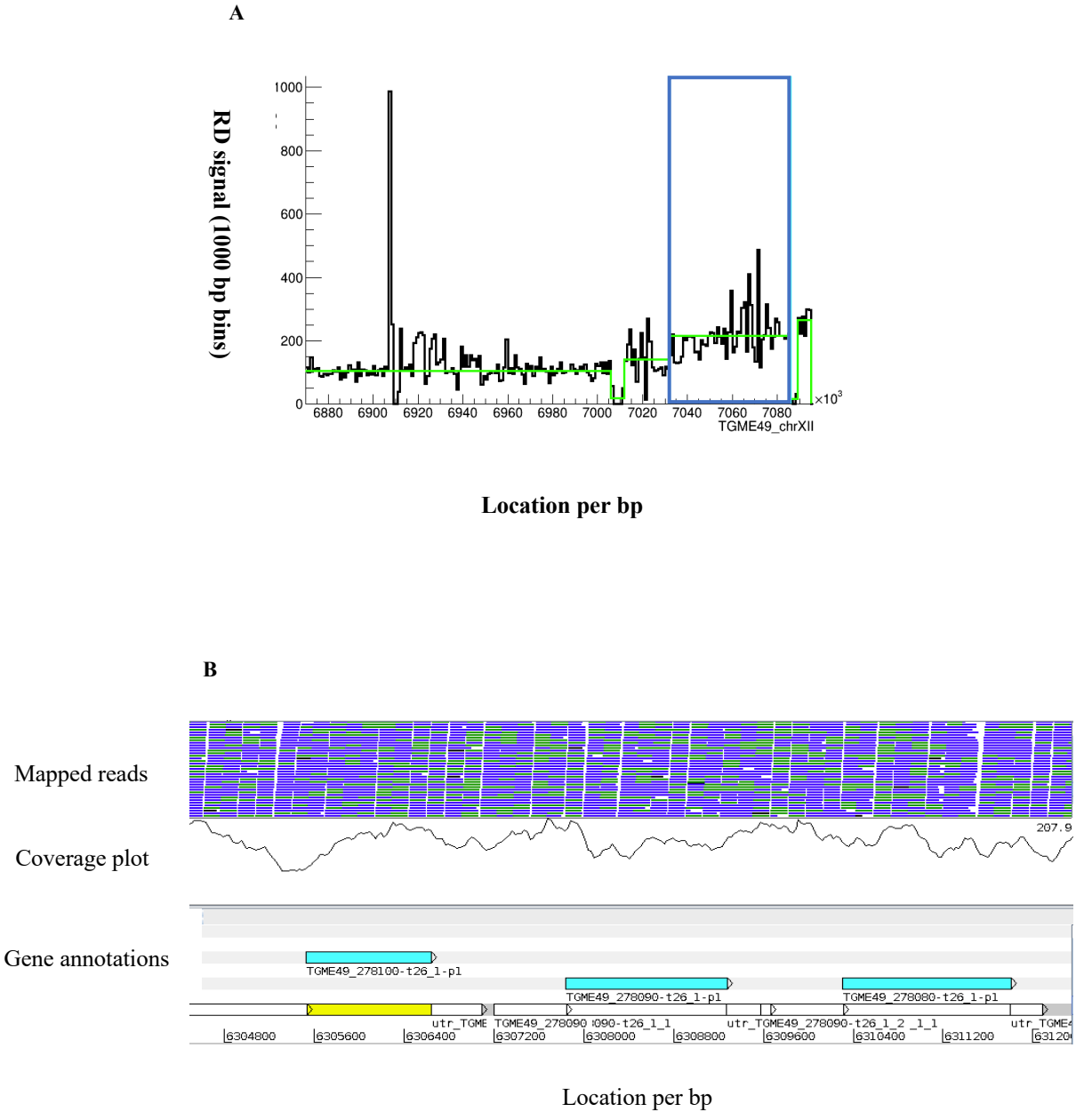



**Figure 5.23**: **A)** The genomic view of the regions that are predicted to be duplicated in chromosome XII. The y axis is the location per bp and the X axis was the read depth (RD) for predicted CNVs. **B)** A example of one of the region of the genes that have clustered in *Toxoplasma gondii* family A protein (TgFAMA) (TGME49_278080, TGME49_278090, TGME49_278100) are highlighted by boxes and their respective locations on the chromosome XII.

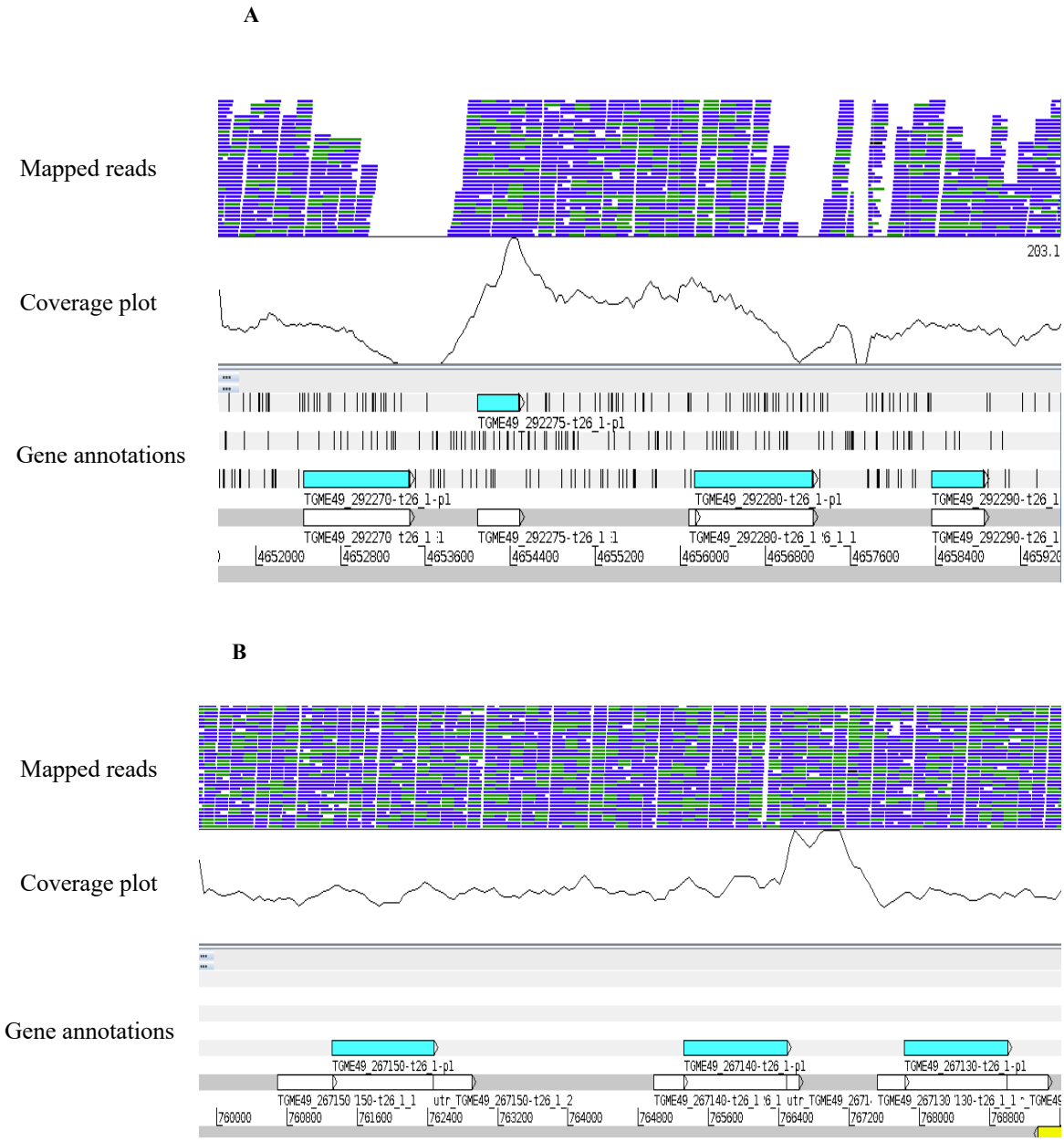



**Figure 5.24**: **A)** The genomic view of the regions that have predicted to be duplicated in chromosome IX from in *T. GT1*. The coverage graph underneath the mapped reads in black includes an examples of enriched gene family SRS includes 3 members of SRS including SRS36C, SRS36E and SRS36D (TGME49_292270- TGME49_2922775- TGME49_292280). **B)** Further example of duplicated genes named as TGME49_267150 (SRS38C), TGME49_267140 (SRS38B) and TGME49_267130 (SRS38A) that were all paralogues of on another.

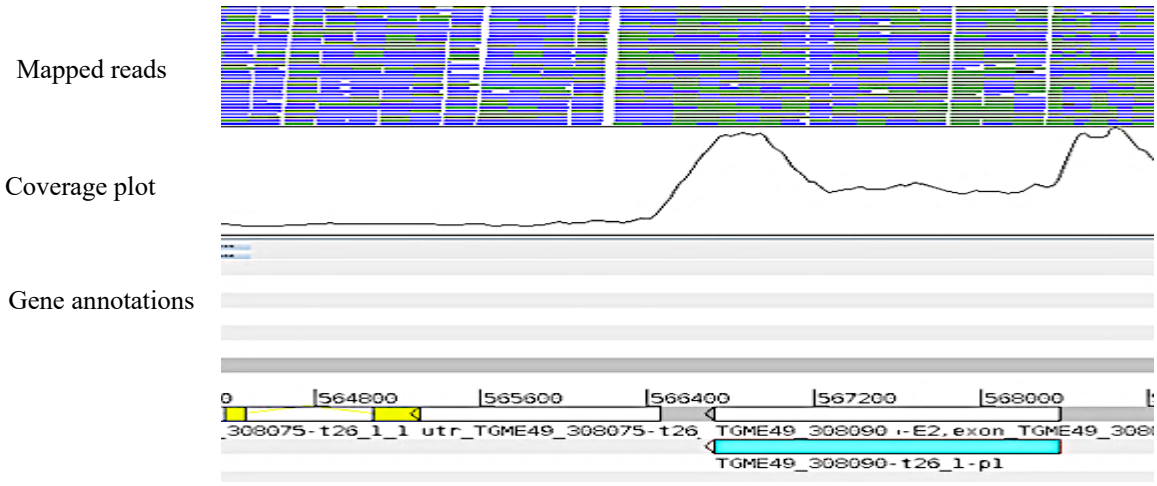



**Figure 5.25: A)** Another example of the genomic view of the regions that are predicted to be duplicated in chromosome XII that have ROP5 in all the strain (TGME49_308090) from 566,721 - 568,370 clustered into different numbers of copies depending on the strain types (I, II and III).

#### 5.3.7.1 GO term analysis of the genes with CNV in *T. gondii* strains

We went on to perform GO term analysis to investigate any significant trend of the duplicated genes that might be involved in host-parasite interaction, host range and other phenotypic traits that shaped the genetic diversity between *T. gondii* strains. As highlighted earlier, the vast majority of annotated genes were hypothetical proteins in all the strains of *T. gondii* and as no GO terms were assignable to these predicted proteins, they have been precluded from the analyses below.

By examining the Molecular Function GO terms assigned to the predicted proteins in the six strains we noticed an enrichment of terms such as protein and DNA binding (Figure 5.26). More particularly, the terms for protein kinase activity, catalytic activity and ATP binding. This suggested that group were gene families members from SRS, ROPs, KRUF and TgFAMs. In addition to this, the biological processes were mainly categorized into several general process. However, the most frequently GO term was protein phosphorylation which may suggest a role of those members in antigenic variation and virulence based on the enrichment terms between strains. Furthermore, we found that the membrane terms were widely observed in all the strains of *T. gondii*. (Figures 5.26-28).

In summary, members of pathogen-specific gene families were expanded in different strain types across the 14 chromosomes. More specifically, we found these duplications encoded in regions towards the ends of the chromosomes, which are regions that shaped the overall genomic architecture of *T. gondii* strains. This has confirmed our hypothesis that the underlying genomic regions that may be responsible for antigenic variations.

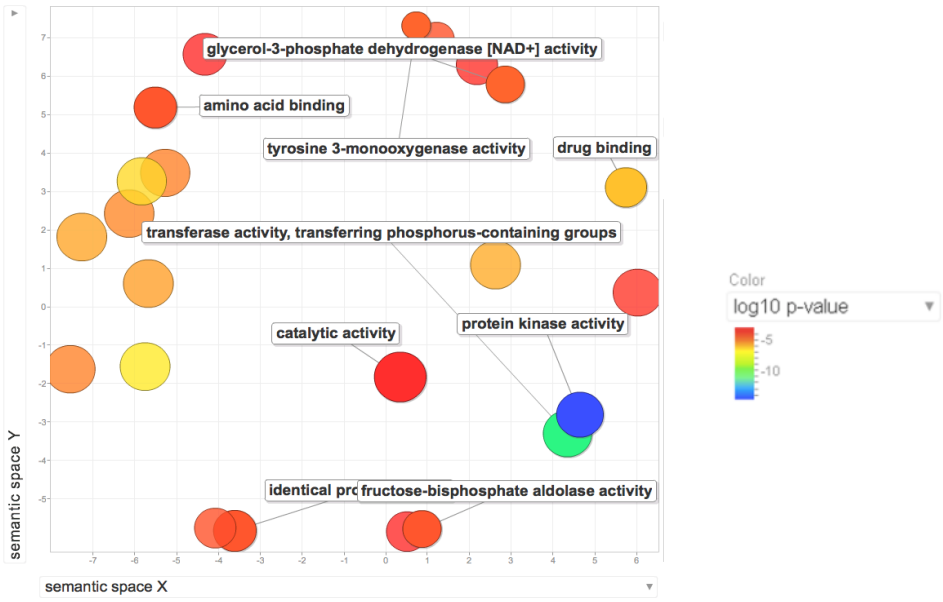



**Figure 5.26**: REVIGO's Scatterplot showing different cluster representatives of molecular function ontologies of pathways of duplicated genes with CNVs in the strains of *T. gondii* after the redundancy reduction were summarised from different strains. The cluster with the large blue and green circle have grater P-values assigned with more significance based on the size of bubbles of more general GO terms. Protein kinase activity terms were found in all *T. gondii* strains. The legend shows the frequencies of the GO term.

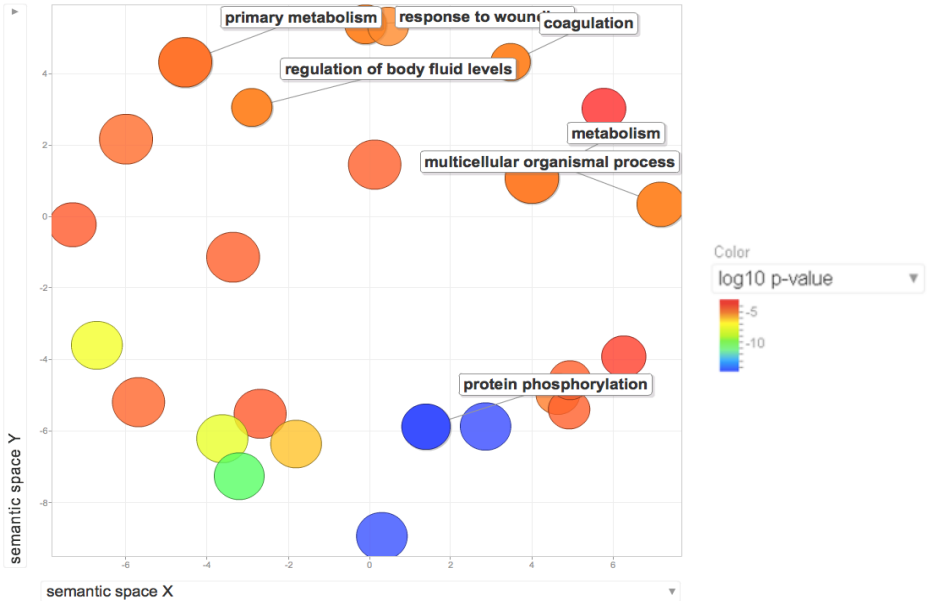



**Figure 5.27**: REVIGO's Scatterplot showing different cluster representatives of biological process ontologies of pathways of duplicated genes with CNVs in the strains of *T. gondii* after the redundancy reduction were summarised from different strains. The clusters with larger blue and green circles have grater P-values assigned with more significance based on the size of bubbles of more general GO terms. Protein phosphorylation process was the highest frequency term in all strains. The legend shows the frequencies of the GO term.

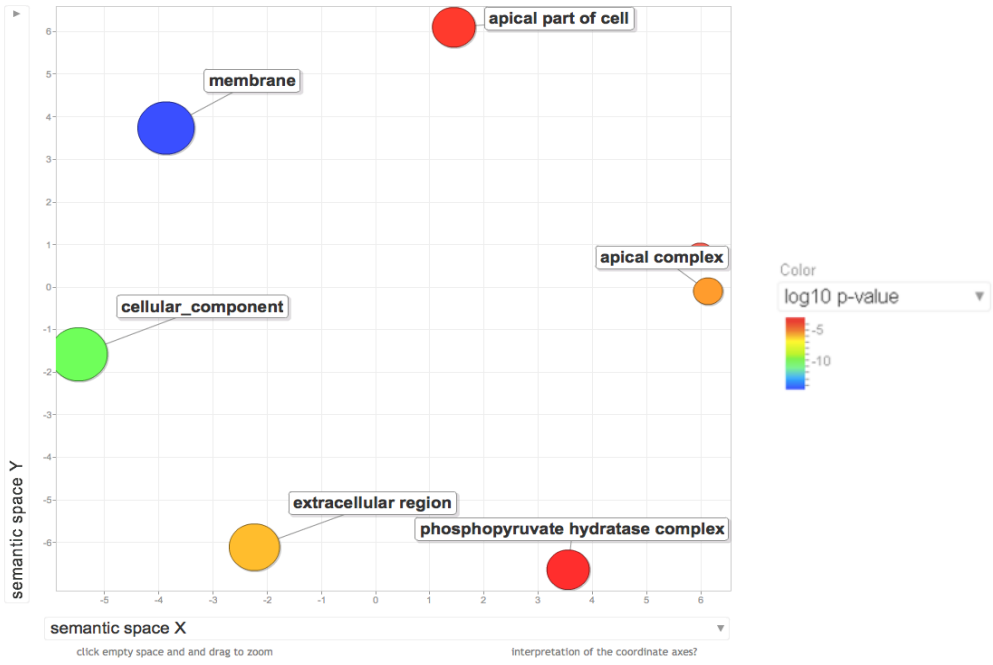



**Figure 5.28**: REVIGO's Scatterplot showing different cluster representatives of cellular component ontologies of pathways of duplicated genes with CNVs in the strains of *T. gondii* after the redundancy reduction were summarised from different strains respectively. The clusters with larger blue and green circles have grater P-values assigned with more significance based on the size of bubbles of more general GO terms. Membrane was the highest frequency among GO terms noticed in all *T. gondii* strains. The legend shows frequencies of the GO term.

### 5.3.8 *De novo* assembly analysis of unmapped reads

We successfully extracted all the reads from the *T.* gondii isolates (*T. MAS; T. P89; T. CAST; T. VEG* and *T. COUG*) that did not map to the ME49 reference genome. In addition, we included a WGS *T. gondii* reference genome ME49 taken from https://www.ncbi.nlm.nih.gov/sra/?term=SRR6793863 to get a more general view of whether there was novel gene content not found in the reference genome with no significant hits or which might be missing from our refence genome during the alignment process using the short read tool as described in Chapter 2. We found significant variation in the proportions of unmapped reads. A total of 35.27%, 20.29%, 16.15 %, 15.8 %, 8.22% and 15.6% were identified in *T. COUG, T. MAS, T. CAST, T. P89, T. VEG* and T. GME49 (taken from SR6793863) respectively; we found no unmapped reads in the strain *T. GT1* (see Table 5.2). The taxonomic analysis was carried out using the MetaPhlAn software and this showed strong signals of contaminations from two different sources. The first source indicated bacterial contamination, namely *Mycoplasma* species. To explore the contents of the unmapped reads furthe , we aligned all the unmapped reads against *Mycoplasma* and *T. gondii* reference genomes using Bowtie2. The mapping stat showed that there were still a large number of reads that did not map to either *T. gondii* or *M. hyorhinis* genomes.

We successfully extracted all the reads that did not map to either species and performed BLAST searches against the NCBI nt database. We found that > 95% of the reads mapped to the African green monkey from which the Vero cell lines were derived. We next extracted all the clean reads and assembled those reads following the same protocol as detailed in the Materials and Methods. Based on the quality assessments for the varied genome assemblies there still appeared to be contamination in those files as indicated by the GC% content per sample. We used BlobTools to identify true biological sequences derived from only the specific target genomes free from contaminations that belonged only to the specific strains of *T. gondii*. The assembled reads were entered into the pipeline and run individually per strain; the unmapped reads were subjected to a series of taxonomic annotations steps that primarily used the GC content of sequences. Based on these inputs, the NCBI

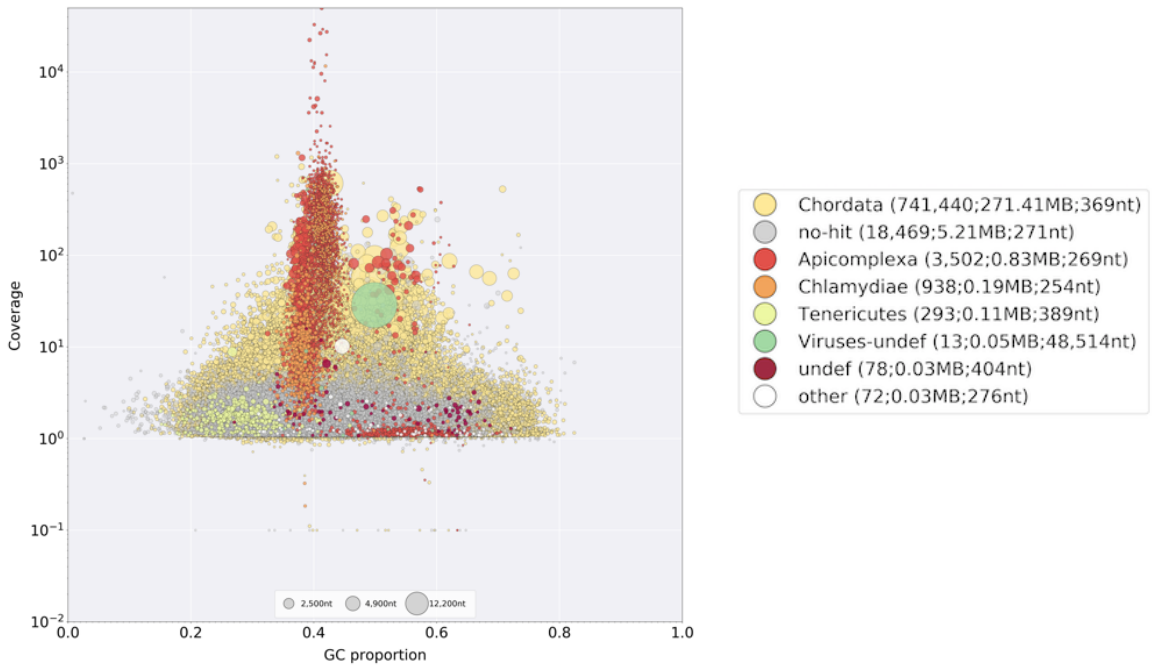
taxonomies were assigned for each sequence in the assembly per strain based on the highest scoring NCBI Taxa ID revealed per sample (Figures 5.29-33).
Not surprisingly, the vast majority of the scaffolds entered were assigned to the phylum of Chordata as we expected earlier from the main source of contamination. The second largest taxonomic group identified in the annotated scaffolds belonged to bacterial species likely caused by experimental contamination. The Blobplots analysis allowed us to identify those reads/assembled contigs free of contamination ready for the next downstream analysis.

We next extracted all the contigs likely to represent true Apicomplexan sequences for all the samples individually. The comparison of the resulting assemblies of the Apicomplexa contigs are summarized in Table 5.8. The *de novo* assemblies totalled 18006, 1928, 1961, 429 and 1596 contigs in *T. MAS, T. CAST, T. P89, T. VEG* and *T. COUG*, respectively. We found a maximum contig size in T. COUG, of 33,126 bp with an N50 of 1512 nucleotides, followed by the *T. P89* strain. There was a significant reduction in the number of contigs in *T. VEG* with 429 contigs and a total length 199,962 bp. We next went on to analyse the gene content of the assemblies to see whether novel genes might explain the intra-species divergence. To achieve this, we annotated the contigs using the Companion pipeline. There was no significant evidence of identified genes in *T. VEG* due to the short reads lengths as we expected from the *de novo* assembly statistics presented in Table 5.8. In the remaining strains, we identified 18 genes, sixteen of them were coding genes includes 3, 7, 4 and 4 in *T. MAS, T. CAST, T. P89* and *T. COUG*, respectively. Two non-coding genes were observed in *T. MAS* and *T. P89* with one gene each (Table 5.9). To further evaluate whether the gene sequences were really novel or actually from the reference genome of *T. gondii* strain ME49, we blasted all the sequences against the NCBI nt database. 14 out of the 16 protein coding genes were annotated as hypothetical proteins with unknown functions in all strains of *T. gondii.* We confirmed that most of the genes had homologues in the other strains, including the reference genome. We thus speculate that the reads were not mapping due to poor mapping with the short reads alignment tool and/or sequencing errors. Interestingly, no significant hits were observed for the two remaining hypothetical proteins. We were, however, able to confirmed that those sequences had homologues in *N. caninum* strain Liverpool with a total length of 134 and 136 amino acids in *T. P89* and *T. COUG* strains respectively.

**Table 5.8:** The overview of assembly statistics of the unmapped reads of the 5 strains of *T. gondii*. *T. GT1* strain was not include due to the smallest percentage of unmapped reads.

| **Assembly statistics metrics** | ***T. MAS*** | ***T. CAST*** | ***T. P89*** | ***T. VEG*** | ***T. COUG*** |
|---|---|---|---|---|---|
| **Number of contigs** | 8006 | 1928 | 1961 | 429 | 1596 |
| **Total bases (bp)** | 1,205,832 | 562,619 | 1,112,176 | 199,962 | 559,742 |
| **Shortest (bp)** | 78 | 78 | 250 | 250 | 78 |
| **Longest (bp)** | 4343 | 4429 | 7478 | 2824 | 33126 |
| **Average length** | 150.6 | 291.,8 | 576.1 | 466.1 | 350.7 |
| **Average GC%** | 42% | 44.8% | 40.5% | 38.8% | 45% |
| **N50** | 1223 | 1281 | 1116 | 1038 | 1512 |
| **N75** | 798 | 805 | 814 | 656 | 796 |
| **L50** | 65 | 41 | 193 | 28 | 30 |
| **L75** | 126 | 82 | 372 | 55 | 77 |

A




B

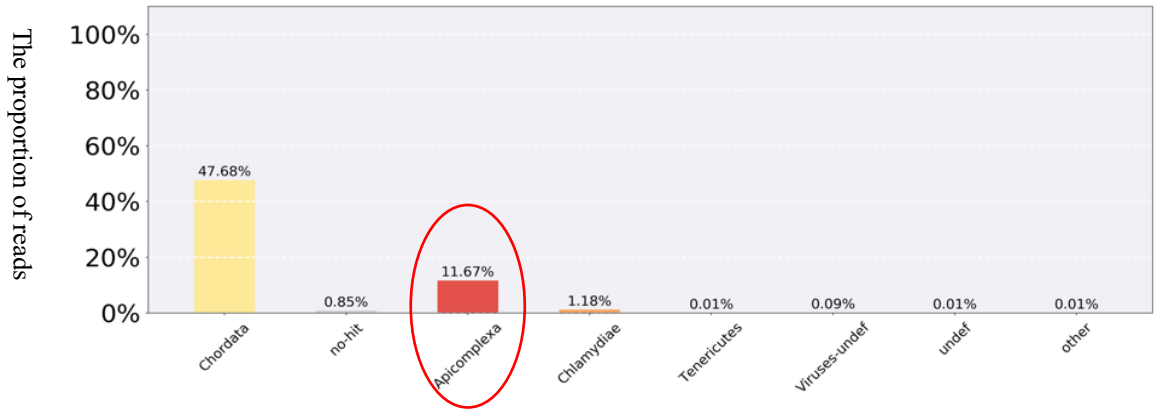
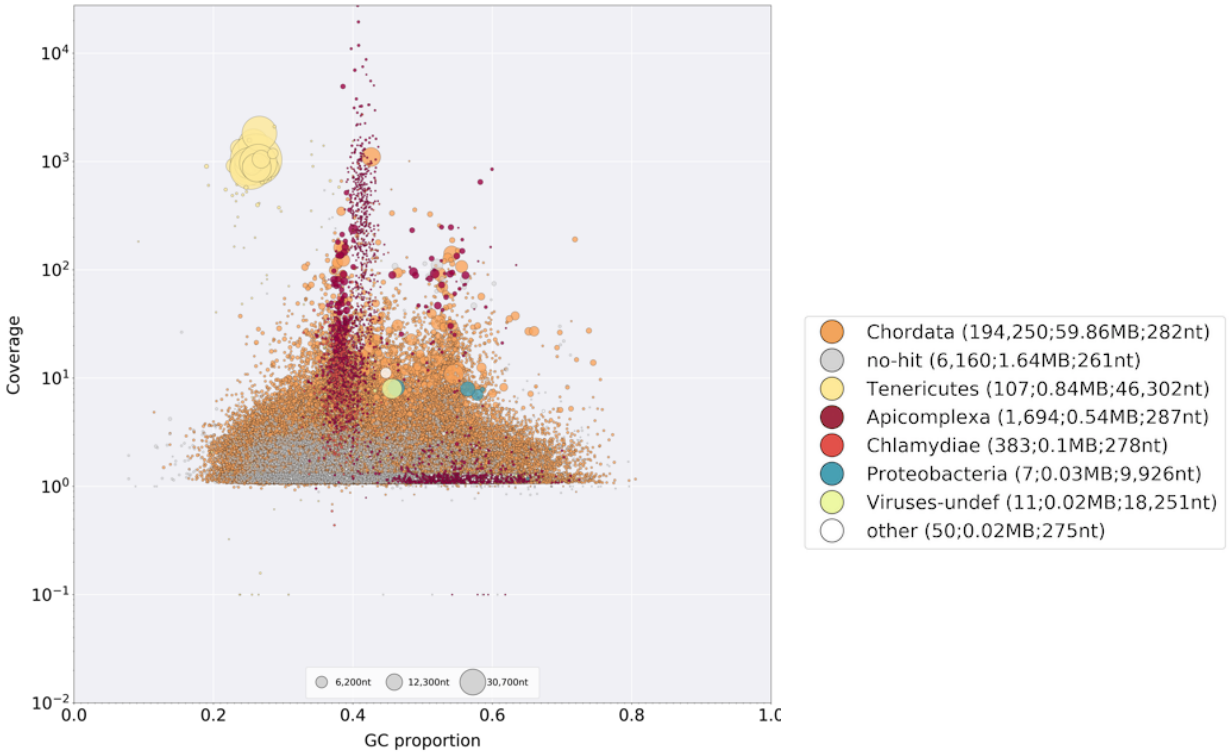



**Figure 5.29: A)** BlobPlot of the assembly done by BlobTools depicted as coloured circles. The legend shows ranks based on the taxonomic order from BLASTN similarity research. The coloured circles were positioned on the X-axis based on their GC proportion and on the Y-axis is the read coverage of the contigs for the *T. MAS* sample. **B)** ReadCovPlot of mapped reads were shown by taxonomic group, the red circle indicates to the Apicomplexa sequences (11.6%).

A




**B**

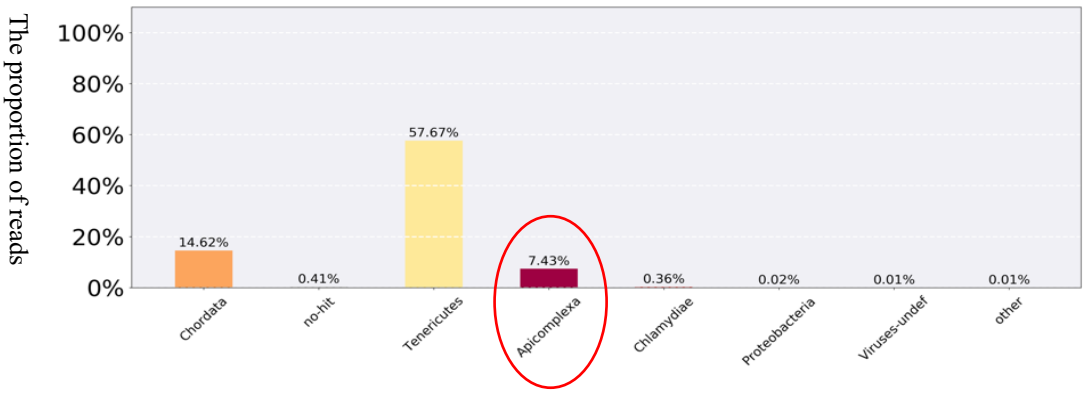
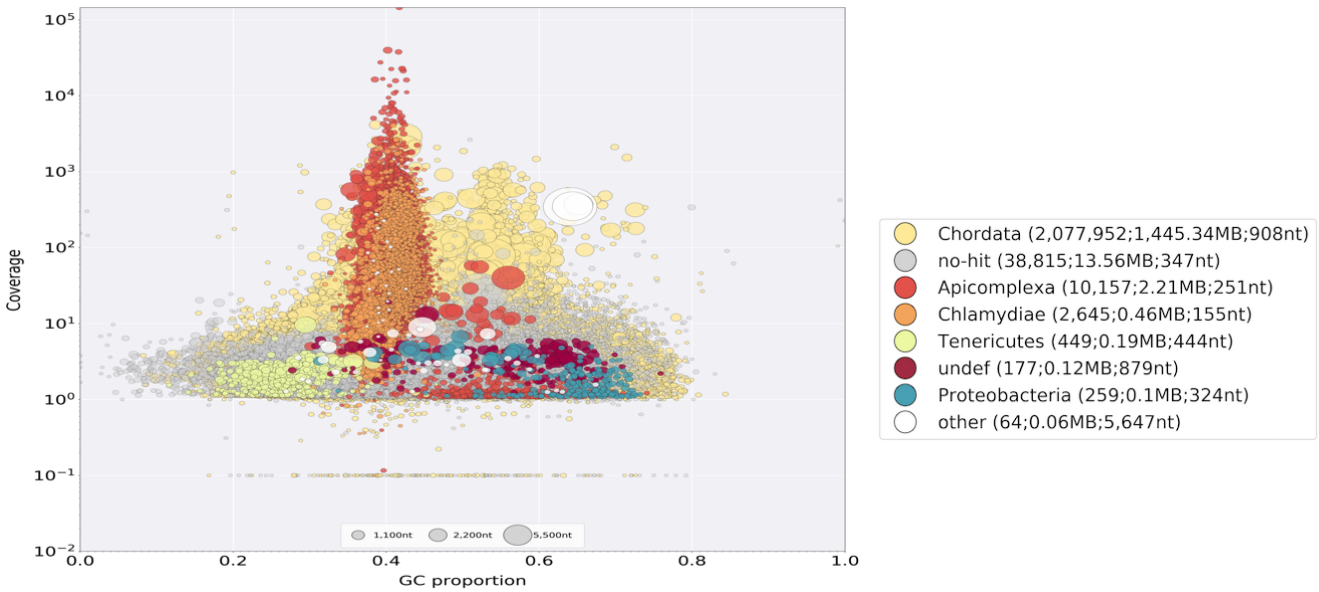



**Figure 5.30: A)** BlobPlot of the assembly done by BlobTools depicted as coloured circles. The legend shows ranks based on the taxonomic order from BLASTN similarity research. The coloured circles were positioned on the X-axis based on their GC proportion and on the Y-axis is the read coverage of the contigs for the *T. CAST* sample. **B)** ReadCovPlot of mapped reads shown by taxonomic group, the red circle indicated to the Apicomplexa sequences (7.4%).

**A**




**B**

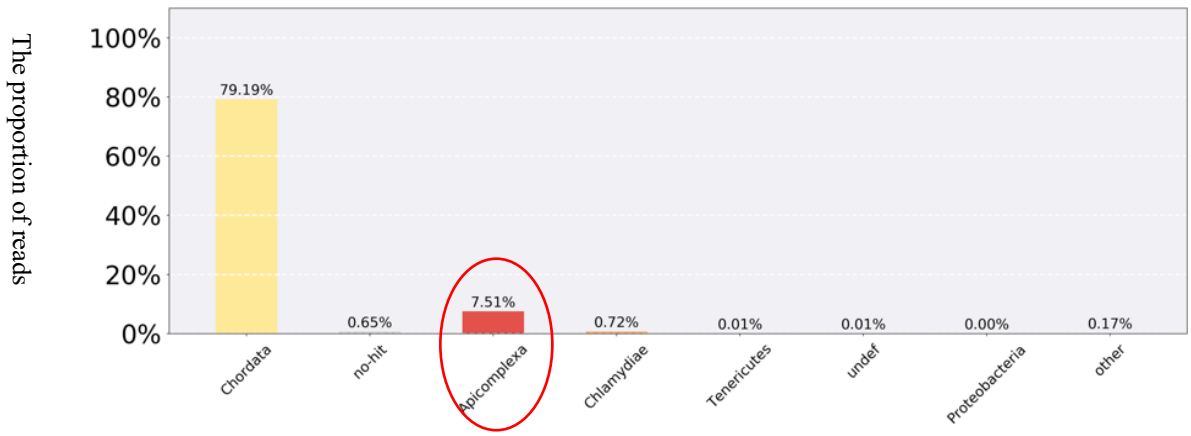
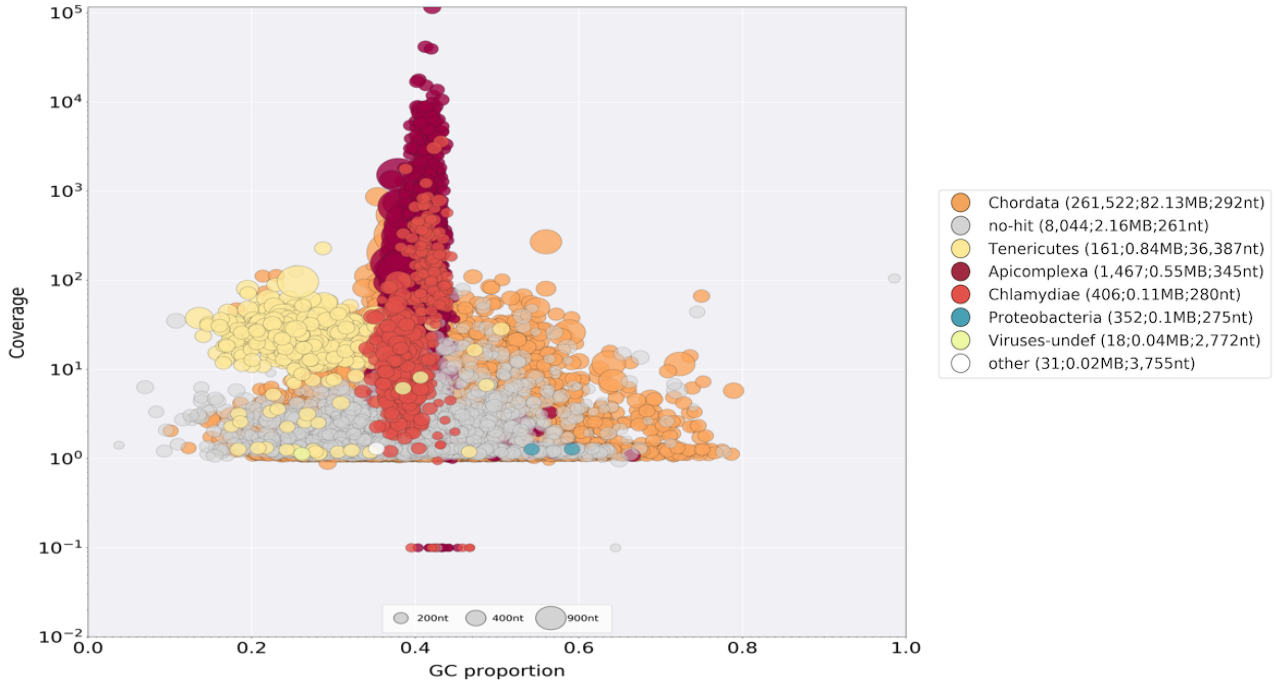



**Figure 5.31: A)** BlobPlot of the assembly done by BlobTools depicted as coloured circles. The legend shows ranks based on the taxonomic order from BLASTN similarity research. The coloured circles were positioned on the X-axis based on their GC proportion and on the Y-axis is the read coverage of the contigs for the *T. P89* sample. **B)** ReadCovPlot of mapped reads shown by taxonomic group, the red circle indicated to the Apicomplexa sequences (7.5%).

**A**




**B**

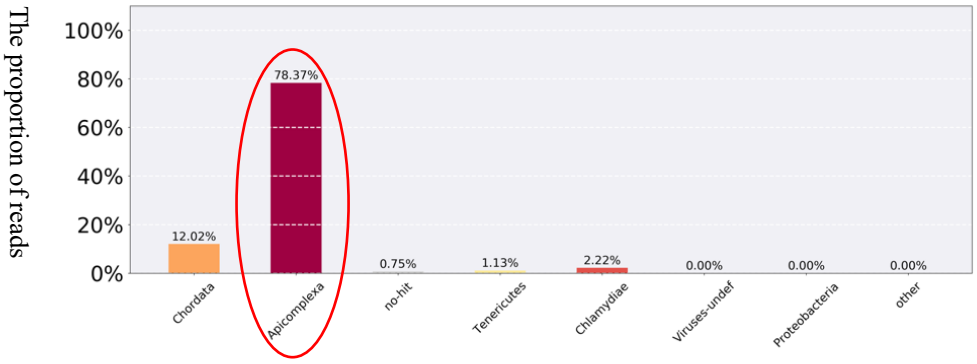
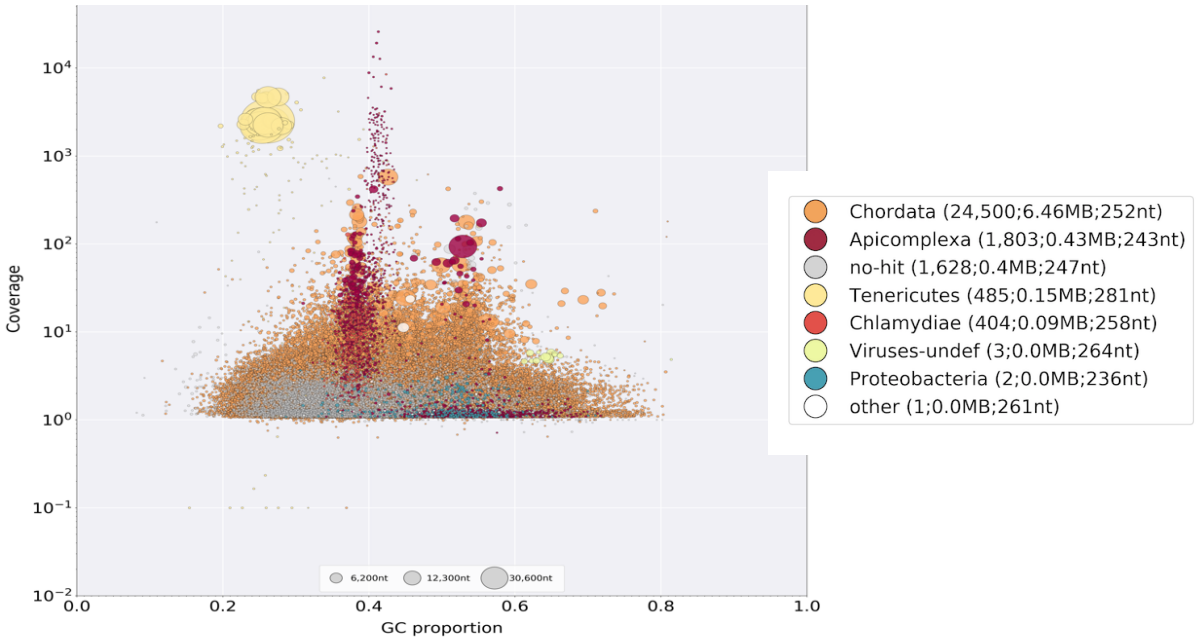



**Figure 5.32: A)** BlobPlot of the assembly done by BlobTools depicted as coloured circles. The legend shows ranks based on the taxonomic order from BLASTN similarity research. The coloured circles were positioned on the X-axis based on their GC proportion and on the Y-axis is the read coverage of the contigs for the *T. COUG* sample. **B)** ReadCovPlot of mapped reads shown by taxonomic group, the red circle indicated to the Apicomplexa sequences (79%).

**A**




**B**

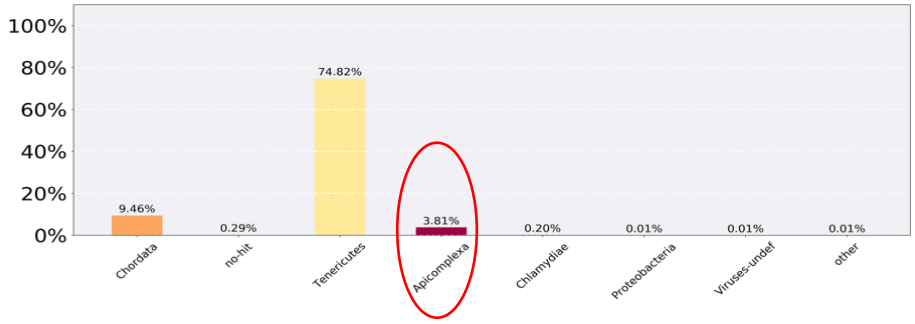



**Figure 5.33: A)** BlobPlot of the assembly done by BlobTools depicted as coloured circles. The legend shows ranks based on the taxonomic order from BLASTN similarity research. The coloured circles were positioned on the X-axis based on their GC proportion and on the Y-axis is the read coverage of the contigs for the *T. VEG* sample. **B)** ReadCovPlot of mapped reads shown by taxonomic group, the red circle indicated to the Apicomplexa sequences (3.8%).

**Table 5.9:** The gene metrics of the annotated sequences belonging to different strains of the *T. gondii* by using Companion tool (https://companion.sanger.ac.uk).

| Genetic metrics | *T. MAS* | *T. CAST* | *T. P89* | *T. COUG* |
|---|---|---|---|---|
| **Number of annotated regions/sequences** | 1 | 1 | 1 | 1 |
| **Number of genes** | 3 | 4 | 7 | 4 |
| **Gene density (genes/mega base)** | 4.18 | 7.1 | 5.4 | 7.1 |
| **Number of coding genes** | 2 | 4 | 6 | 4 |
| **Number of pseudogenes** | 0 | 0 | 0 | 1 |
| **Number of genes with function** | 1 | 1 | 1 | 2 |
| **Number of pseudogenes with function** | 0 | 0 | 0 | 1 |
| **Number of non-coding genes** | 1 | 0 | 1 | 0 |
| **Number of genes with multiple CDSs** | 0 | 0 | 0 | 0 |
| **Overall GC%** | 42.41% | 44.82% | 40.5 | 44.99% |
| **Coding GC%** | 51.83% | 64.01% | 60.66% | 62.53% |

## 5.4 Discussions

Here we have presented a comprehensive analysis of the genomes of *T. gondii* strains collected from different regions in the world from varied hosts using whole genome sequencing (WGS). Our analysis compared those resequenced isolates against the reference genome *T. gondii* strain ME49. The findings reported here suggest that there was substantial shared gene content, as expected, and that this shared content represented all general biological processes and functions. However, our data also clearly show that genetic diversity between the strains exists, notably in the expansion of the pathogen- specific gene families includes SAG-related sequence (SRS), rhoptry protein (ROPs), micronemes (MICs), dense granules (GRAs), *Toxoplasma gondii* gene families (TgFAMs) and Lysine-Arginine rich Unidentified Function (KRUF) that might underlie other phenotypic differences of specific strains. These findings are consistent with those of (Bontell *et al.*, 2009; Reid *et al.*, 2012; Walzer and Boyle, 2012; Adomako-Ankomah *et al.*, 2014; Lorenzi *et al.*, 2016) who confirmed that there was a significant association between these specific gene families and pathogenesis, host range, adaption and evolution.

To our knowledge, this is the first SNP annotation study based on NGS of multiple strains of *T. gondii* that set out to determine the functional consequences of those SNPs in the virulence-associated genes. Cheng *et a.,*(2015) compared the genetic variation of two Chinese strains that belonging to the *T. GT1* strain type I, identifying a total of 505,856 and 505,654 SNPs in Wh3 and Wh6 respectively with the majority located in exonic regions, totalling 285 and 56 high impact SNPs in Wh3 and Wh6 strains respectively. A total of 26 genes contained variations; most of those belonged to known gene families, including SRSs, ROPs and GRAs. More significantly, RON3 and GRA3 show high levels of expression involved in virulence. Both strains shared two polymorphic effectors includes ROP16 and GRA15 genes with type I, II and III isolates (Cheng *et al.*, 2015). Our results further support the idea of antigenic variation by expansion in a large number of gene families that are involved in host-parasite interactions and cover different aspects of host preference such as transmission route, life styles and other phenotypic characteristics. This in particular relates to the diversity in telomeric regions that enable high rates of recombination as seen in different species, e.g. VARs genes in *Plasmodium falciparum* (Kyes, Kraemer and

Smith, 2007; Petter and Duffy, 2015) and VSGs genes in *Trypanosoma brucei* (Berriman *et al.*, 2005; Jackson *et al.*, 2012).This pattern of gene family expansion and contractions suggest that this phenomenon might also explain the diversity not only between species but also between the distinct strains of *T. gondii* (Sibley and Boothroyd, 1992), (Lehmann *et al.*, 2000), (Su *et al.*, 2012), (Shwab *et al.*, 2014).

Our data also report for the first time the number of the high impact SNPs located in the multiple gene families involved in pathogenesis causing deleterious effects on the protein products. The vast majority of those SNPs were located in the three most abundant families namely SRSs, ROPs and TgFAMs. More importantly, they occurred in members of the ROP family including the ROP5, ROP16 and ROP18 genes that potentially play a key role in strain variations in virulence. Such high impact mutations will likely affect the functions of those genes, especially in highly virulent strains such as *T. GT1*(Ong, Reese and Boothroyd, 2010; Behnke *et al.*, 2011; Khan, Shaik, *et al.*, 2014).

In the present study we also compared CNVs from distinct strains to highlight what could be the functional consequences of duplicated genes; we were particularly interested in ascertaining whether those genes were enriched for specific gene families. Cheng *et al*, 2015 identified were 85 CNV and 90 duplication in the Wh3 and Wh6 isolates respectively that were generally in the exonic regions, which is a higher number of CNVs than in our study. This result may be explained by the small bin size used in their analysis, which allowed for a higher number of CNV duplications. Our findings are in line with other previous investigations that confirmed most of the CNVs were tandemly duplicated and belonged to two main gene family in *T. gondii* (SRS and TgFAMs).

One of the other findings to emerge from this analysis is that despite the low quantities of genomic DNA, samples could still be processed into sequencing libraries and generate data suited to these types of diversity studies. Those finding are in keeping with previous studies, which reported that despite low coverage of sequencing data SNPs could still be reliably called. We demonstrated that the low coverage for those samples was due to large amount of contaminating reads, most likely originating from contaminates such as host and bacterial genomes as a consequences of tissue culture

experiments. In accordance with the present findings, previous studies have highlighted frequent contamination with *Mycoplasma* species which are hard to detect due to the small size of the bacteria. At the time of culturing the *Toxoplasma* isolates the tissue culture facility used was not routinely screened for the presence of *Mycoplasma*. This is clearly a key omission as the presence of *Mycoplasma* is known to have an impact on the host -pathogen mechanisms and growth rate of the parasites. More importantly, the bacteria can causing aberrations in the chromosomes, which will later influence on the final output of the genomic analysis by changing the functions of the genes that are mainly used in this type of the cell lines (Koboldt *et al.*, 2010; Gouin *et al.*, 2015; Peng *et al.*, 2015; Usman *et al.*, 2017).

Hence, it could conceivably be hypothesised that the removal of large amount of reads will affect the size of the strain-specific genomic sequences. In addition to this, several factors could explain this finding such as the long period of time required for a extensive number of cell passaging in culture conditions since first date of isolation that might be change the nature of the parasites and alter some phenotypic traits such as growth rate, transmigration and virulence of the tachyzoites. More specifically in the highly virulent strain *T. GT1*(type I) and intermediate virulence strain *T. VEG* (type III) (Morrissette and Sibley, 2002; M.-L. Dardé, 2004b; L. David Sibley *et al.*, 2009; Khan and Grigg, 2017).

# Chapter 6:

## 6.1 General discussion

This thesis has presented a deeper insight into the genomes of two closely related coccidian pathogens *Toxoplasma gondii* and *Neospora caninum.* This work has expanded on previous knowledge gleaned from comparative genomic analysis of the *T. gondii* and *N. caninum* reference genomes and extended into population-level analyses. This study has been one of the first attempts to thoroughly examine the sequences of distinct isolates of *T. gondii* and *N. caninum* using next generation sequencing strategies to expand on our understanding of how the variation driving the difference in gene families may contribute to isolate pathogenesis, host restriction, transmission routes, disease manifestations and other phenotypic traits (Dubey, Lindsay and Speer, 1998; Dubey, Schares and Ortega-Mora, 2007a; Reid *et al.*, 2012; Adomako-Ankomah *et al.*, 2014; Ramaprasad *et al.*, 2015; Lau *et al.*, 2016; Lorenzi *et al.*, 2016).

One well-known comparative genomic and transcriptomic sequencing study that is often cited in research on the differences between the two apicomplexan organisms is that of Reid (2012), who found that the two genomes were highly syntenic due to a large number of shared conserved regions with more than 90% of one to one orthologues in the protein coding genes between the two parasites (Reid *et al.*, 2012; DeBarry and Kissinger, 2014; Reid, 2015). One of the more significant outputs to emerge from this previous comparison is the role of expanded gene families, in particular the surface antigen gene family) in *N. caninum*, that might play a role in the host restriction of this parasite and which might answer the questions as to why *N. caninum* has a limited number of hosts compared to *T. gondii* and what is the relationship between expansions of different gene families and the phenotypic differences between the two parasites. The work in Chapter 3 reviewed comparative genomic analyses between the *T. gondii* and *N. caninum* reference genome assemblies. We systematically re-examined the list of species-specific genes and multiple gene families (SRSs, ROPs, MICs, GRAs, TSF and KRUFs) that were published previously by Reid *et al.,* (2012).

Our data are in agreement with Reid's (2012) findings which showed that there was a significant increase in the number of *T. gondii*- specific genes rather than *N. caninum*-specific genes. However, we were able to add novel insights to the differences in the expansion of the gene families involved in host -parasite interactions between the two studies. A possible explanation for this difference might be that the annotation (both structural and functional) of these initial reference genomes has been updated since the first release. Reid et al. (2012) used the gene models of *T. gondii* and *N. caninum* as released via ToxoDB version 5.2, which had fewer protein coding genes than the most recent release of the *Toxoplasma* genome (ToxoDB version 29) which was used in the present study. This current release of ToxoDBv29 had more genes annotated that accounted for 18% and 2% in *T. gondii* (ME49) and *N. caninum* (Liverpool), respectively.

The different methods, algorithms and query parameters used will have resulted in different gene candidate lists after combining results from the earlier and more recent versions of the genome assemblies. Hence, the lack of comprehensive annotations of the *T. gondii* and *N. caninum* genomes will likely have resulted in missing annotations, more specifically for those large number of hypothetical proteins with unknown functions that might be part of species-specific genes and members of known or novel gene families. This does indeed highlight the value of re-examining such comparative analyses as new information and interpretations are added over time.

We will release our updates via the genome databases such that our efforts will help improve the overall gene annotations as part of community-driven efforts to ensure good quality resources for the further analyses. Our findings are in keeping with previous genomic and transcriptomic sequencing projects which confirmed that there is strong evidence of improvement in the gene annotations by identifying novel genes that are not available in the public domain or correcting those gene models. This is particularly important in our continued effort to understands the structure of the *T. gondii and N. caninum* genomes, more significantly, to discover the potential pathogenic effectors that under selective pressure have covered different aspects of diversity, in particular in virulence , host range, defence strategy and life style (Behnke *et al.*, 2011; Reid *et al.*, 2012; Adomako-Ankomah *et al.*, 2014; Walzer *et al.*, 2014; Krishna *et al.*, 2015; Ramaprasad *et al.*, 2015; Lau *et al.*, 2016).

This study also offered additional evidence from our comparisons that the differences between these two closely related genomes was related to the large number of pathogenic gene families that invariably lead to loss of synteny in the genomes and which play a key role in antigenic variations, host parasite mechanisms and phenotypic variations gene (Gardner *et al.*, 2002; Barry *et al.*, 2003; Su *et al.*, 2003; Berriman *et al.*, 2005; Pain *et al.*, 2005, 2008; Okamoto and McFadden, 2008; Debarry and Kissinger, 2011; Wasmuth *et al.*, 2012).

Here we successfully identified further proteins that were added to the previously described gene families from the three apical organelles includes rhoptry (ROPs), micronemes (MICs) and dense granules (GRAs) that make up a large number of protein coding genes in the *T. gondii* genome, and we also provide more members of SAG1-related sequences known the Surface Antigen gene family (SRSs) in *N. caninum*. The most interesting finding was the significant expansion of one specific gene family known as *Toxoplasma gondii* family proteins (TgFAMs) in *T. gondii*. In particular of interest, the TgFAMC subfamily was expanded further in *T. gondi* in telomeric clustering. The old name for this specific *T. gondii* family was used in the Reid study as *Toxoplasma* specific family (TSF) with fewer genes than we identified in our current study. Further research should be undertaken to investigate the functions of this family in *T. gondii* and *N. caninum* due to a greater number of family members in *T. gondii.* This would allow us to address key questions such as what the consequences of the absence the TgFAMC gene family from the *N. caninum* genome? Is this family involved in host range restriction? Does this family mediate other phenotypic changes that must contribute to virulence, dynamics of transmission and/or host -parasite interactions?

Previous reports by Pollard *et al.,* (2008) and Reid *et al.*, (2012) showed that another parasitic gene family named SAG-Unrelated Surface Antigens (SUSA) was expanded on chromosomes VI and XII in both genomes. Our findings presented here suggest that the SUSA genes previously identified in *N. caninum* had orthologues with TgFAMA gene family that are highly expanded in the *T. gondii* genome encoded at the same chromosomal location. It seems possible that the SUSA-1 and SUSA-2 genes noticed in *N. caninum* are homologues to TgFAMs due to recent gene annotations revealing further evidence of expansion in more gene families during the process of

gene annotation improvements (Pollard *et al.*, 2008; Adomako-Ankomah *et al.*, 2014; Dalmasso *et al.*, 2014; Lorenzi *et al.*, 2016). We then expanded on our current understanding of *Toxoplasma* and *Neospora* genomics by investigating the population diversity within each of the parasitic species.

To our knowledge, this is the first whole genomic sequencing project that was undertaken with three phenotypically distinct strains of *N. caninum named NC-Liverpool*, *NC-1* and *NC-Bahia* using NGS sequencing techniques. Limited investigations have been performed previously to determine the level of polymorphism in this species. Calarco *et al*., (2018) first examined the SNPs in *N. caninum* by comparing the transcriptomic data of two different strains, *NC-Liverpool* and *NC-Nowra*, detecting a total of 3,130 SNPs and 6,123 indels with evidence of clustering of polymorphisms on chromosomes XI and XII. In Chapter 4, we used a wide range of bioinformatic pipelines to determine the genetic diversity among those three distinct strains of *N. caninum* to generate different SNP data sets and compared them to the published reference genome. We analysed the proportions of high impact SNPs that might have an impact on the function of the proteins in hotspot regions that might be controlling several biological functions such as host - parasite interactions.

We were able to carry out these analyses despite only being able to generate limited data for some of the isolates, in particular *NC-Bahia*. We were only able to generate limited amount of input DNA for the sequencing libraries for this isolate which could be attributed to two possible explanations. Firstly, it was isolated from a 1999 culture from infected brain (Gondim *et al.*, 2001). Since then, a high number of passages have been performed and this caused lower growth rate over time, impacting the viability and the density of the cell culture (Ammerman, 2009). Another factor that might explain the limited amount of DNA is cell culture contamination, specifically, bacterial *Mycoplasma* contamination, which commonly has an effect on cell proliferation and alters the growth rate. Given their small size, *Mycoplasma* cells are difficult to detect in routinely cell culture passages and are highly resistance to the antibiotics in the medium used. A second source of contamination was from the host genome, the African Green Monkey Vero cells used as host for the tachyzoites.

Using the recent reference genome of *N. caninum* strain Liverpool, the number of SNPs called from the resequencing Liverpool sample was significantly lower due to the high level of conservation between our cultured *NC-Liverpool* and the published reference genome. Other possible explanation for this lowering level of divergence between the sequences of our cultured sample and the refence genome was the genetic stability of Liverpool isolate from 1998 until now, which is in line with another recent comparative study of the transcriptome of *N. caninum* strains (Calarco, Barratt and Ellis, 2018). These three strains differ significantly in their pathogenicity *in vivo* in mice models (Al-Qassab, Reichel and Ellis, 2010; Dubey and Schares, 2011). By comparing the three strains, there was evidence of higher pathogenicity in the *NC-Liverpoo*l strain compared *to NC-1* and *NC-Bahia* in experimental infection (Lindsay and Dubey, 1989; Schock *et al.*, 2001; Collantes-Fernández *et al.*, 2006). The lowest pathogenicity was noticed in *NC-Bahia* and this confirmed that there was genetic variation between strains and represents pathogenic variability (Chryssafidis *et al.*, 2014). We also identified SNP within the three genomes of *N. caninum*, that can contribute to understanding the differences in phenotype between strains. More significantly, when we compared the nonsynonymous SNPs and the SNPs density across the 14 chromosomes, they might be positively associated with the function of the proteins that have a key role in virulence between strains.

We are confident in the robustness of our analysis pipelines. We used the Genome Analysis Toolkit (GATK) for variant discovery. This software package is routinely used in the field and should mitigate against platform artefacts (potential GC biases, sequencing error) including efficiency of the library preparation method given the low input DNA and variant caller sensitivities and specificities that were previously reported (Minoche, Dohm and Himmelbauer, 2011; Chen *et al.*, 2013; Nascimento *et al.*, 2016).

In Chapter 5, we presented a comprehensive analysis of *T. gondii* strains collected from different regions in the world and from different hosts. Our analysis compared those resequencing isolates against the reference genome *T. gondii* strain ME49. The findings reported here suggest that there was a high level of synteny due to sharing large proportions of the core *T. gondii* genome. We report here for the first time the number of high impact SNPs located in the multiple gene families which are involved

in pathogenesis causing deleterious effects on the protein products. The vast majorities of those variations were located in the three most abundant families including SRS, ROP and TgFAM. More importantly, in members of the ROP family including ROP5, ROP16 and ROP18 that potentially play a key role in strain variations in virulence. Such high impact SNPs in those genes will affect the functions of those genes .Comparison of the findings with those of other studies confirmed the hypothesis of selective pressure which increases the fitness of *T. gondii* due to the high rate of polymorphism between strains during genetic evolution specifically in these gene members that were directly involved in host parasite interactions (Paterson, Vogwill, Buckling, Benmayor, Andrew J. Spiers, *et al.*, 2010; Kerstes *et al.*, 2012; Auld and Tinsley, 2015). A further possible explanation for the divergence between strains might be down to the geographical distributions of *T. gondii* strains in North/Central/South America and Europe. In addition to this, migration of the definitive host such as cat and intermediate hosts such as humans and other hosts between countries might have contributed some novel history of the host genome as we expected (Shwab *et al.*, 2014).

In addition to investigating single nucleotide polymorphisms, we also assessed the impact of copy number variations. These findings broadly support the work of other studies in this area linking copy number variations with duplicated genes that were involved in the pathogenesis; this might also answer the questions that we asked in our aims; are there correlations between the CNVs and the patterns of clustering in specific locations across the different genomes of *T. gondii* and *N. caninum*? We hence compared those CNV from distinct strains to highlight what could be the functional consequences of those duplicated genes, more specifically whether those genes were enriched for some specific gene families or clustered within a specific genomic location. Cheng *et al*, (2015) reported that < 90 duplications in *two* Toxoplasma isolates (Wh3 and Wh6) were generally in the exonic regions. This presented a higher number of CNV compared to our study which may be explained by the fact that a small CNV region was used in their analysis. Our results are in line with some previous investigations that confirmed that most CNVs represent tandemly duplicated SRS and TgFAMs gene family members which play a significant role in antigenic variations (Kissinger and DeBarry, 2011; DeBarry and Kissinger, 2014; Cheng *et al.*, 2015; Reid, 2015; Lorenzi *et al.*, 2016).

In the present study, we determined the polymorphisms of 6 *T. gondii* strains collected from different geographical locations and isolated from varied host origins to examine the genetic diversity between them. Our high-resolution genome-wide SNP comparisons clearly show that some of these strains have a high number of SNPs. More significantly, nonsynonymous SNPs were noticed in two strains; *T. CAST* and *T. GT1*. In agreement with previous findings, type I, which the strain *T. GT1* belongs to had higher virulence and was associated with the highest pathogenicity in humans while types II and III were less virulent. In addition, further evidence of the high number of SNPs in recombinant strains that did not belonging to the three previously clonal types defined (I, II and III) uncovered new aspects of population structure of *T. gondii* and present new pathogenic effectors to guide us to understand different phenotypes. Aspects such as immune response of the host, migration, pathogenesis, geographical distribution, virulence and evolution can be tested as additional complete genomes of *T. gondii* are sequenced (Lehmann *et al.*, 2000; Kim and Weiss, 2004; Saeij, Boyle and Boothroyd, 2005; Khan Asis *et al.*, 2011; Minot *et al.*, 2012).

Aside from looking at polymorphisms and CNVs, we also investigated the data that did not map to the respective reference genomes. Such unmapped reads might encode novel genes or missing genes from our references genomes. In *N. caninum*, we confirmed that there are some sequences that encode predicted genes which might be involved in genetic diversity between the isolates included in this study (Gouin *et al.*, 2015; Whitacre *et al.*, 2015; Usman *et al.*, 2017). Most of the genes identified were hypothetical proteins with unknown functions from the small contigs assembled from the *NC-Bahia* and *NC-Liverpool* samples, which had orthologues to *T. gondii* and other apicomplexan species such as *Hammondia hammondi* but were not in the sequences of the *N. caninum* Liverpool genome. This reflected that there are novel gene contents in the *NC-Bahia* not found in *NC-Liverpool* genome. In *NC-Bahia* sequence, four novel genes appeared including two hypothetical proteins, one containing a predicted Kelch motif/Galactose oxidase, central domain and a fourth containing a GYF domain with orthologues in other *T. gondii* strains. However, there were hits of significant similarity to two hypothetical proteins with 84 -100% identity to the *NC-Liverpool* genome. This suggested that in fact those two gene models were indeed found in the *N. caninum* reference genome assembly but that, likely owing to mis-mapping reads, they appeared to be isolate-specific novel content.

In our resequencing *NC-Liverpool* sample, we identified one gene encoding a protein of unknown function that was missing from the reference genome. Interestingly, three genes assigned to new proteins included hypothetical proteins and a SNARE domain containing protein in *NC-Liverpool*, with putative orthologues to *T. gondii*. In the unmapped reads of the *T. gondii* isolates, we found a number of 'novel' genes that encode proteins of unknown function some of which had a putative homologue in *N. caninum* and other apicomplexan species. However, no homology existed between the two further hypothetical proteins in the *T. P89* and the *T. COUG* and the *T. gondii* reference genomes. Again, this highlighted the importance of assessing coding potential in unmapped reads.

## 6.2 Concluding remarks and contributions to the field

In conclusion, this current study use first detailed further study to investigate the range of variations between *N. caninum* and *T. gondii* strains, confirmed that there were differences between the two apicomplexan species which could drive the variations in host preference, virulence, mechanism of infection, transmission route, and geographical distributions between strains of *N. caninum* and *T. gondii*. Returning to the question posed at the beginning of this project, it is now possible to state that the two species differ greatly in the expansion of enriched gene families that are often located in the telomeric regions in both genomes. Additional findings supported the hypothesis of gene family expansions in SRSs genes in *N. caninum* genome and TgFAMs in *T. gondii*. This work has hence contributed to improving the gene annotations of the *N. caninum* and *T. gondii* genomes and those improvements will be made available via the genome database resource. Overall, these results provide opportunities for further analyses of the population structures of both species by adding NGS data from a number of isolates that are drawn from a wide range of hosts into the public domain.

### 6.3 Future work and scientific contributions achievements

Further experiments will be required to investigate the impact that the genomic variations reported in this present study have on our understanding of the biology of *N. caninum* and *T. gondii*, in particular to determine the role that the pathogenic genes play in pathogenesis, defence strategies and host restriction. Further studies regarding the role of advanced sequencing techniques would be worthwhile by using a broader range of DNA and RNA samples from additional strains for both parasites.

In addition, it will be important to perform knockout protocols to test the functions of the genes identified especially, for studying the role of genes which have been sequenced but whose functions have not been determined. The genomic and transcriptomic resources could shed more light on the basic differences between the two genomes such as determining specific time points during infections and gene expression during host -parasite- interactions. Future research on proteomic experiments might extend the explanations of variations in the protein expression and the functions among different stages of the *T. gondii* and *N. caninum* and other apicomplexan parasites life cycles.

In further investigations, the use of further proteomic data will provide additional evidences to discover new genes that have evidence for protein level expression from different proteomic experiments which can be added to the current genome annotations more importantly the virulence genes that involved in host parasite interaction, adaption mechanism in different habitat, survival strategy and host ranges which potentially reveal the biological differences between the *T. gondii* and *N. caninum* parasites.

Another future research is needed to confirm our novel findings by performing further genome editing technologies named clustered regularly interspaced short palindromic repeats (CRISPR-Cas9) to test the functions of essential genes that are involved in pathogenesis under the tested conditions from different strains in both parasites. In addition, CRISPR-Cas9 might prove an important area for preventing and treatment of toxoplasmosis and neosporosis diseases by identifying the targeted genes that play a key role in host -parasites interaction and signalling metabolic pathways.

## References


Abnizova, I., Boekhorst, R. te and Orlov, Y. L. (2017) 'Computational Errors and Biases in Short Read Next Generation Sequencing', *Journal of Proteomics & Bioinformatics*, 10(1), pp. 1–17. doi: 10.4172/jpb.1000420.

Adomako-Ankomah, Y. *et al.* (2014) 'Differential locus expansion distinguishes Toxoplasmatinae species and closely related strains of Toxoplasma gondii', *mBio*, 5(1). doi: 10.1128/mBio.01003-13.

Ajay, S. S. *et al.* (2011) 'Accurate and comprehensive sequencing of personal genomes', *Genome Research*. doi: 10.1101/gr.123638.111.

Ajzenberg, D. *et al.* (2004) 'Genetic diversity, clonality and sexuality in Toxoplasma gondii', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2004.06.007.

Al-Bajalan, M. M. M. *et al.* (2017) 'Toxoplasma gondii and Neospora caninum induce different host cell responses at proteome-wide phosphorylation events; a step forward for uncovering the biological differences between these closely related parasites', *Parasitology Research*., 116(10), pp. 2707–2719. doi: 10.1007/s00436-017-5579-7.

Al-Qassab, S. *et al.* (2009) 'Genetic diversity amongst isolates of Neospora caninum, and the development of a multiplex assay for the detection of distinct strains', *Molecular and Cellular Probes*, 23(3–4), pp. 132–139. doi: 10.1016/j.mcp.2009.01.006.

Al-Qassab, S. E., Reichel, M. P. and Ellis, J. T. (2010) 'On the biological and genetic diversity in Neospora caninum', *Diversity*. doi: 10.3390/d2030411.

Ammerman (2009) 'Growth and Maintenance of Vero Cell Line', *Curr Protoc Microbiol*. doi: 10.1002/9780471729259.mca04es11.

Anderson, M. L. *et al.* (1992) 'Neospora-like protozoan infection as a cause of abortion in dairy cattle', *Journal of Veterinary Diagnostic Investigation*. doi: 10.1177/104063879200400228.

Anderson, M. W. and Schrijver, I. (2010) 'Next generation DNA sequencing and the future of genomic medicine', *Genes*. doi: 10.3390/genes1010038.

Ari, Ş. and Arikan, M. (2016) 'Next-Generation Sequencing: Advantages, Disadvantages, and Future', in *Plant Omics: Trends and Applications*. doi: 10.1007/978-3-319-31703-8_5.

Ashley, E. A., Pyae Phyo, A. and Woodrow, C. J. (2018) 'Malaria', *The Lancet*, 391(10130), pp. 1608–1621. doi: 10.1016/S0140-6736(18)30324-6.

Atkinson, R. *et al.* (1999) 'Comparison of the biological characteristics of two isolates of Neospora caninum', *Parasitology*. doi: 10.1017/S0031182098003898.

Auld, S. K. and Tinsley, M. C. (2015) 'The evolutionary ecology of complex lifecycle parasites: Linking phenomena with mechanisms', *Heredity*, 114(2), pp. 125–132. doi: 10.1038/hdy.2014.84.

Barber, J. S. and Trees, A. J. (1998) 'Naturally occurring vertical transmission of Neospora caninum in dogs', in *International Journal for Parasitology*, pp. 57–64. doi: 10.1016/S0020-7519(97)00171-9.

Barragan, A. and Sibley, L. D. (2003) 'Migration of Toxoplasma gondii across biological barriers', *Trends in Microbiology*, pp. 426–430. doi: 10.1016/S0966-842X(03)00205-1.

Barry, J. D. *et al.* (2003) 'Why are parasite contingency genes often associated with telomeres?', *International Journal for Parasitology*. doi: 10.1016/S0020-7519(02)00247-3.

Barry, J. D. *et al.* (2005) 'What the genome sequence is revealing about trypanosome antigenic variation.', *Biochemical Society transactions*. doi: 10.1042/BST20050986.

Bauer, D. C. (2011) 'Variant calling comparison', *Nature precedings*. doi: 10.1038/npre.2011.6107.1.

Behnke, M. S. *et al.* (2011) 'Virulence differences in Toxoplasma mediated by amplification of a family of polymorphic pseudokinases', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1015338108.

Behnke, M. S. *et al.* (2014) 'Toxoplasma gondii merozoite gene expression analysis with comparison to the life cycle discloses a unique expression state during enteric development', *BMC Genomics*. doi: 10.1186/1471-2164-15-350.

Behnke, M. S. *et al.* (2015) 'Rhoptry Proteins ROP5 and ROP18 Are Major Murine Virulence Factors in Genetically Divergent South American Strains of Toxoplasma gondii', *PLoS Genetics*, 11(8). doi: 10.1371/journal.pgen.1005434.

Berriman, M. *et al.* (2005) 'The genome of the African trypanosome Trypanosoma brucei', *Science*. doi: 10.1126/science.1112642.

Bezerra, M. A. *et al.* (2017) 'Constitutive expression and characterization of a surface SRS (NcSRS67) protein of Neospora caninum with no orthologue in Toxoplasma gondii', *Parasitology International*., 66(2), pp. 173–180. doi: 10.1016/j.parint.2017.01.010.

Bontell, I. L. *et al.* (2009) 'Whole genome sequencing of a natural recombinant Toxoplasma gondii strain reveals chromosome sorting and local allelic variants', *Genome Biology*, 10(5). doi: 10.1186/gb-2009-10-5-r53.

Boothroyd, J. C. (2009) 'Expansion of host range as a driving force in the evolution of Toxoplasma', *Memorias do Instituto Oswaldo Cruz*. doi: 10.1590/S0074-02762009000200009.

Boothroyd, J. C. and Grigg, M. E. (2002) 'Population biology of Toxoplasma gondii and its relevance to human infection: Do different strains cause different disease?', *Current Opinion in Microbiology*. doi: 10.1016/S1369-5274(02)00349-1.

Boyle, J. P. *et al.* (2006) 'Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like Toxoplasma gondii.', *Proceedings of the National Academy of Sciences of the United States of America*, 103(27), pp. 10514–10519. doi: 10.1073/pnas.0510319103.

Bradley, P. J. *et al.* (2005) 'Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in Toxoplasma gondii', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M504158200.

Buermans, H. P. J. and Den Dunnen, J. T. (2014) 'Next generation sequencing technology: Advances and applications ☆', *BBA - Molecular Basis of Disease*, 1842, pp. 1932–1941. doi: 10.1016/j.bbadis.2014.06.015.

Butcher, B. A. *et al.* (2011) 'Toxoplasma gondii rhoptry kinase rop16 activates stat3 and stat6 resulting in cytokine inhibition and arginase-1-dependent growth control', *PLoS Pathogens*. doi: 10.1371/journal.ppat.1002236.

Buxton, D. *et al.* (1989) 'Trial of a novel experimental Toxoplasma iscom vaccine in pregnant sheep', *British Veterinary Journal*, 145(5), pp. 451–457. doi: 10.1016/0007-1935(89)90053-5.

Calarco, L., Barratt, J. and Ellis, J. (2018) 'Genome Wide Identification of Mutational Hotspots in the Apicomplexan Parasite Neospora caninum and the Implications for Virulence', *Genome Biology and Evolution*, 10(9), pp. 2417–2431. doi: 10.1093/gbe/evy188.

Camejo, A. *et al.* (2014) 'Identification of three novel Toxoplasma gondii rhoptry proteins', *International Journal for Parasitology*, 44(2). doi: 10.1016/j.ijpara.2013.08.002.

Carruthers, V. B. (2002) 'Host cell invasion by the opportunistic pathogen Toxoplasma gondii', *Acta Tropica*. doi: 10.1016/S0001-706X(01)00201-7.

Carruthers, V. B. and Sibley, L. D. (1997) 'Sequential protein secretion from three distinct organelles of Toxoplasma gondii accompanies invasion of human fibroblasts.', *European journal of cell biology*, 73(2), pp. 114–23. doi: 10.7320/FlMedit23.223.

Carruthers, V. and Boothroyd, J. C. (2007) 'Pulling together: an integrated model of Toxoplasma cell invasion', *Current Opinion in Microbiology*. doi: 10.1016/j.mib.2006.06.017.

Cesbron-Delauw, M. F. (1994) 'Dense-granule organelles of Toxoplasma gondii: Their role in the host-parasite relationship', *Parasitology Today*, pp. 293–296. doi: 10.1016/0169-4758(94)90078-7.

Chen, J. *et al.* (2014) 'Toxoplasma gondii: Protective immunity induced by rhoptry protein 9 (TgROP9) against acute toxoplasmosis', *Experimental Parasitology*, 139(1), pp. 42–48. doi: 10.1016/j.exppara.2014.02.016.

Chen, L. F. *et al.* (2018) 'Comparative studies of Toxoplasma gondii transcriptomes: Insights into stage conversion based on gene expression profiling and alternative splicing', *Parasites and Vectors*. doi: 10.1186/s13071-018-2983-5.

Chen, Y. C. *et al.* (2013) 'Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly', *PLoS ONE*. doi: 10.1371/journal.pone.0062856.

Cheng, W. *et al.* (2015) 'Variation detection based on next-generation sequencing of type Chinese 1 strains of Toxoplasma gondii with different virulence from China', *BMC Genomics*. BMC Genomics, 16(1), pp. 1–9. doi: 10.1186/s12864-015-2106-z.

Chryssafidis, A. L. *et al.* (2014) 'Pathogenicity of Nc-Bahia and Nc-1 strains of Neospora caninum in experimentally infected cows and buffaloes in early pregnancy', *Parasitology Research*. doi: 10.1007/s00436-014-3796-x.

Church, D. M. *et al.* (2015) 'Extending reference assembly models', *Genome Biology*, 16(1), pp. 2–6. doi: 10.1186/s13059-015-0587-3.

Cingolani, P. *et al.* (2012) 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff', *Fly*. doi: 10.4161/fly.19695.

Clough, B. and Frickel, E. M. (2017) 'The Toxoplasma Parasitophorous Vacuole: An Evolving Host–Parasite Frontier', *Trends in Parasitology*. doi: 10.1016/j.pt.2017.02.007.

Collantes-Fernández, E. *et al.* (2006) 'Comparison of Neospora caninum distribution, parasite loads and lesions between epidemic and endemic bovine abortion cases', *Veterinary Parasitology*. doi: 10.1016/j.vetpar.2006.05.030.

Coppens, I., Sinai, A. P. and Joiner, K. A. (2000) 'Toxoplasma gondii exploits host low-density lipoprotein receptor- mediated endocytosis for cholesterol acquisition', *Journal of Cell Biology*, 149(1), pp. 167–180. doi: 10.1083/jcb.149.1.167.

Cuypers, B. *et al.* (2017) 'Genome-wide SNP analysis reveals distinct origins of Trypanosoma evansi and Trypanosoma equiperdum', *Genome Biology and Evolution*. doi: 10.1093/gbe/evx102.

Dalmasso, M. C. *et al.* (2014) 'Characterization of Toxoplasma gondii subtelomeric-like regions: Identification of a long-range compositional bias that is also associated with gene-poor regions', *BMC Genomics*. doi: 10.1186/1471-2164-15-21.

Dardé, M.-L. (2004b) 'Genetic analysis of the diversity in Toxoplasma gondii.', *Annali dell'Istituto superiore di sanità*, 40(1), pp. 57–63. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3533945&tool=pmcentrez&rendertype=abstract.

Dardé, M. L. *et al.* (1992) 'Isoenzyme Analysis of 35 Toxoplasma gondii Isolates and the Biological and Epidemiological Implications', *The Journal of Parasitology*. doi: 10.2307/3283305.

Dardé, M. L. (2004) 'Genetic analysis of the diversity in Toxoplasma gondii', *Annali dell'Istituto Superiore di Sanita*.

Dardé, M. L., Ajzenberg, D. and Su, C. (2013) 'Molecular Epidemiology and Population Structure of Toxoplasma gondii', in *Toxoplasma Gondii: The Model Apicomplexan - Perspectives and Methods: Second Edition*. doi: 10.1016/B978-0-12-396481-6.00003-9.

Darde, M. L., Bouteille, B. and Pestre-Alexandre, M. (1988) 'Isoenzymic characterization of seven strains of Toxoplasma gondii by isoelectrofocusing in polyacrylamide gels', *American Journal of Tropical Medicine and Hygiene*. doi: 10.4269/ajtmh.1988.39.551.

Sibley, D. L. *et al.* (2009) 'Genetic diversity of Toxoplasma gondii in animals and humans', *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi: 10.1098/rstb.2009.0087.

Sibley, D. L. (2011) 'Invasion and intracellular survival by protozoan parasites', *Immunological Reviews*. doi: 10.1111/j.1600-065X.2010.00990.x.

Debarry, J. D. and Kissinger, J. C. (2011) 'Jumbled genomes: Missing apicomplexan synteny', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msr103.

DeBarry, J. D. and Kissinger, J. C. (2014) 'A survey of innovation through duplication in the reduced genomes of twelve parasites', *PLoS ONE*. doi: 10.1371/journal.pone.0099213.

Depristo, M. A. *et al.* (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*. doi: 10.1038/ng.806.

Donahoe, S. L. *et al.* (2015) 'A review of neosporosis and pathologic findings of Neospora caninum infection in wildlife', *International Journal for Parasitology: Parasites and Wildlife*. doi: 10.1016/j.ijppaw.2015.04.002.

Dou, J. *et al.* (2012) 'Reference-free SNP calling: Improved accuracy by preventing incorrect calls from repetitive genomic regions', *Biology Direct*, 7, pp. 1–9. doi: 10.1186/1745-6150-7-17.

Drexler, H. G. and Uphoff, C. C. (2002) 'Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention', in *Cytotechnology*. doi: 10.1023/A:1022913015916.

Dubey, J. P. *et al.* (1988) 'Newly recognized fatal protozoan disease of dogs', *J. Am. Vet. Med. Assoc.*

Dubey, J. P. *et al.* (1998) 'Effect of gamma irradiation on unsporulated and sporulated Toxoplasma gondii oocysts', *International Journal for Parasitology*. doi: 10.1016/S0020-7519(97)83432-7.

Dubey, J. P. (1998) '*Toxoplasma gondii* oocyst survival under defined temperatures', *The Journal of Parasitology*. doi: 10.2307/3284606.

Dubey, J. P. (2005) 'Unexpected oocyst shedding by cats fed Toxoplasma gondii tachyzoites: In vivo stage conversion and strain variation', *Veterinary Parasitology*. doi: 10.1016/j.vetpar.2005.06.007.

Dubey, J. P. (2006) 'Comparative infectivity of oocysts and bradyzoites of Toxoplasma gondii for intermediate (mice) and definitive (cats) hosts', *Veterinary Parasitology*, 140(1–2), pp. 69–75. doi: 10.1016/j.vetpar.2006.03.018.

Dubey, J. P. *et al.* (2011) 'Genetic characterisation of Toxoplasma gondii in wildlife from North America revealed widespread and high prevalence of the fourth clonal type', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2011.06.005.

Dubey, J. P. (2013) 'The History and Life Cycle of Toxoplasma gondii', in *Toxoplasma Gondii: The Model Apicomplexan - Perspectives and Methods: Second Edition*. doi: 10.1016/B978-0-12-396481-6.00001-5.

Dubey, J. P. and Jones, J. L. (2008) 'Toxoplasma gondii infection in humans and animals in the United States', *International Journal for Parasitology*, pp. 1257–1278. doi: 10.1016/j.ijpara.2008.03.007.

Dubey, J. P., Lindsay, D. S. and Speer, C. A. (1998) 'Structures of Toxoplasma gondii tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts', *Clinical Microbiology Reviews*. doi: PMC106833.

Dubey, J. P. and Schares, G. (2011) 'Neosporosis in animals-The last five years', *Veterinary Parasitology*. doi: 10.1016/j.vetpar.2011.05.031.

Dubey, J. P., Schares, G. and Ortega-Mora, L. M. (2007a) 'Epidemiology and control of neosporosis and Neospora caninum', *Clinical Microbiology Reviews*, pp. 323–367. doi: 10.1128/CMR.00031-06.

Dubey, J. P., Schares, G. and Ortega-Mora, L. M. (2007b) 'Epidemiology and control of neosporosis and Neospora caninum', *Clinical Microbiology Reviews*. doi: 10.1128/CMR.00031-06.

Dubremetz, J. F. and Lebrun, M. (2012) 'Virulence factors of Toxoplasma gondii', *Microbes and Infection*, 14(15), pp. 1403–1410. doi: 10.1016/j.micinf.2012.09.005.

El-Sayed, N. M. *et al.* (2005) 'The genome sequence of Trypanosoma cruzi, etiologic agent of chagas disease', *Science*, 309(5733), pp. 409–416. doi: 10.1126/science.1112631.

Ellegren, H. and Galtier, N. (2016) 'Determinants of genetic diversity', *Nature Reviews Genetics*. doi: 10.1038/nrg.2016.58.

English, E. D., Adomako-Ankomah, Y. and Boyle, J. P. (2015) 'Secreted effectors in Toxoplasma gondii and related species: Determinants of host range and pathogenesis?', *Parasite Immunology*, 37(3), pp. 127–140. doi: 10.1111/pim.12166.

Estrada-Rivadeneyra, D. (2017) 'Sanger sequencing', *FEBS Journal*. doi: 10.1111/febs.14319.

Ferguson, D. J. P. *et al.* (1999) 'In vivo expression and distribution of dense granule protein 7 (GRA7) in the exoenteric (tachyzoite, bradyzoite) and enteric (coccidian) forms of Toxoplasma gondii', *Parasitology*, 119(3), pp. 259–265. doi: 10.1017/S0031182099004692.

Flegr, J. *et al.* (2014) 'Toxoplasmosis - A global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries', *PLoS ONE*, 9(3). doi: 10.1371/journal.pone.0090203.

Forrester, S. J. and Hall, N. (2014) 'The revolution of whole genome sequencing to study parasites', *Molecular and Biochemical Parasitology*, 195(2), pp. 77–81. doi: 10.1016/j.molbiopara.2014.07.008.

Francia, M. E. and Striepen, B. (2014) 'Cell division in apicomplexan parasites', *Nature Reviews Microbiology*. doi: 10.1038/nrmicro3184.

Frazão-Teixeira, E. *et al.* (2011) 'Multi-locus DNA sequencing of Toxoplasma gondii isolated from Brazilian pigs identifies genetically divergent strains', *Veterinary Parasitology*. doi: 10.1016/j.vetpar.2010.09.030.

Frenkel, J. K., Dubey, J. P. and Miller, N. L. (1970) 'Toxoplasma gondii in cats: Fecal stages identified as coccidian oocysts', *Science*. doi: 10.1126/science.167.3919.893.

Fuentes, I. *et al.* (2001) 'Genotypic characterization of Toxoplasma gondii strains associated with human toxoplasmosis in spain: Direct analysis from clinical samples', *Journal of Clinical Microbiology*. doi: 10.1128/JCM.39.4.1566-1570.2001.

Gajria, B. *et al.* (2008) 'ToxoDB: An integrated toxoplasma gondii database resource', *Nucleic Acids Research*. doi: 10.1093/nar/gkm981.

Gardner, M. J. *et al.* (2002) 'Genome sequence of the human malaria parasite Plasmodium falciparum', *Nature*. doi: 10.1038/nature01097.

Gold, D. A. *et al.* (2015) 'The Toxoplasma dense granule proteins GRA17 and GRA23 mediate the movement of small molecules between the host and the parasitophorous vacuole', *Cell Host and Microbe*. doi: 10.1016/j.chom.2015.04.003.

Gondim, L. F. P. *et al.* (2001) 'Isolation of Neospora caninum from the brain of a naturally infected dog, and production of encysted bradyzoites in gerbils', *Veterinary Parasitology*. doi: 10.1016/S0304-4017(01)00493-9.

Gondim, L F P *et al.* (2004) 'TRANSMISSION OF NEOSPORA CANINUM BETWEEN WILD AND DOMESTIC ANIMALS', *J. Parasitol*, 90(6), pp. 1361–1365.

Gondim, Luis F P *et al.* (2004) 'Variation of the Internal Transcribed Spacer 1 Sequence within Individual Strains and among Different Strains of Neospora caninum', *Source: The Journal of Parasitology J. Parasitol*, 90(901), pp. 119–122. Available at: http://www.jstor.org/stable/3286136.

Goodswen, S. J. *et al.* (2015) 'Improving the gene structure annotation of the apicomplexan parasite Neospora caninum fulfils a vital requirement towards an in silico-derived vaccine', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2015.01.006.

Gouin, A. *et al.* (2015) 'Whole-genome re-sequencing of non-model organisms: Lessons from unmapped reads', *Heredity*. doi: 10.1038/hdy.2014.85.

Grigg, M. E. *et al.* (2001a) 'Success and virulence in Toxoplasma as the result of sexual recombination between two distinct ancestries', *Science*. doi: 10.1126/science.1061888.

Grigg, Michael E. *et al.* (2001b) 'Unusual Abundance of Atypical Strains Associated with Human Ocular Toxoplasmosis', *The Journal of Infectious Diseases*. doi: 10.1086/322800.

El Hajj, H. *et al.* (2007) 'ROP18 is a rhoptry kinase controlling the intracellular proliferation of Toxoplasma gondii', *PLoS Pathogens*. doi: 10.1371/journal.ppat.0030014.

Hall, N. and Carlton, J. (2005) 'Comparative genomics of malaria parasites', *Current Opinion in Genetics and Development*. doi: 10.1016/j.gde.2005.09.001.

Hassan, M. A. *et al.* (2012) 'De novo reconstruction of the Toxoplasma gondii transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs', *BMC Genomics*. doi: 10.1186/1471-2164-13-696.

Hedges, S. B. *et al.* (2015) 'Tree of life reveals clock-like speciation and diversification', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msv037.

Hedges, S. B., Dudley, J. and Kumar, S. (2006) 'TimeTree: A public knowledge-base of divergence times among organisms', *Bioinformatics*. doi: 10.1093/bioinformatics/btl505.

Hehl, A. B. *et al.* (2015) 'Asexual expansion of Toxoplasma gondii merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families to attach, invade, and replicate within feline enterocytes.', *BMC genomics*, 16(66). doi: 10.1186/s12864-015-1225-x.

Hemphill, a and Gottstein, B. (1996) 'Identification of a major surface protein on Neospora caninum tachyzoites.', *Parasitology research*.

Hide, G. *et al.* (2009) 'Evidence for high levels of vertical transmission in Toxoplasma gondii', *Parasitology*. doi: 10.1017/S0031182009990941.

Hill, D. and Dubey, J. P. (2002) 'Toxoplasma gondii: Transmission, diagnosis, and prevention', *Clinical Microbiology and Infection*. doi: 10.1046/j.1469-0691.2002.00485.x.

Holmdahl, O. J. M. *et al.* (1995) 'Characterization of the first european isolate of neospora caninum (dubey, carpenter, speer, topper and uggla)', *Parasitology*. doi: 10.1017/S0031182000077039.

Howe, Daniel K *et al.* (1997) 'Determination of Genotypes of Toxoplasma gondii Strains Isolated from Patients with Toxoplasmosis', *JOURNAL OF CLINICAL MICROBIOLOGY*, 35(6), pp. 1411–1414.

Howe, D. K. *et al.* (1998) 'The p29 and p35 immunodominant antigens of Neospora caninum tachyzoites are homologous to the family of surface antigens of Toxoplasma gondii', *Infection and Immunity*.

Howe, D. K. and Sibley, L. D. (1995a) 'Toxoplasma gondii comprises three clonal lineages: Correlation of parasite genotype with human disease', *Journal of Infectious Diseases*, 172(6), pp. 1561–1566. doi: 10.1093/infdis/172.6.1561.

Hunter, C. A. and Sibley, L. D. (2012) 'Modulation of innate immunity by Toxoplasma gondii virulence effectors', *Nature Reviews Microbiology*. doi: 10.1038/nrmicro2858.

Huynh, M. H. *et al.* (2015) 'Structural basis of Toxoplasma gondii MIC2-associated protein interaction with MIC2', *Journal of Biological Chemistry*, 290(3), pp. 1432–1441. doi: 10.1074/jbc.M114.613646.

Jackson, A. P. *et al.* (2012) 'Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1117313109.

Jensen, K. D. C. *et al.* (2015) 'Toxoplasma gondii superinfection and virulence during secondary infection correlate with the exact ROP5/ROP18 allelic combination', *mBio*, 6(2), pp. 1–15. doi: 10.1128/mBio.02280-14.

Jones, J. L. *et al.* (2001) 'Toxoplasma gondii infection in the United States: Seroprevalence and risk factors', *American Journal of Epidemiology*. doi: 10.1093/aje/154.4.357.

Jung, C., Lee, C. Y. F. and Grigg, M. E. (2004) 'The SRS superfamily of Toxoplasma surface proteins', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2003.12.004.

Kemp, L. E., Yamamoto, M. and Soldati-Favre, D. (2013) 'Subversion of host cellular functions by the apicomplexan parasites', *FEMS Microbiology Reviews*, 37(4), pp. 607–631. doi: 10.1111/1574-6976.12013.

Kerstes, N. A. G. *et al.* (2012) 'Antagonistic experimental coevolution with a parasite increases host recombination frequency', *BMC Evolutionary Biology*. doi: 10.1186/1471-2148-12-18.

Khan, A. *et al.* (2005) 'Composite genome map and recombination parameters derived from three archetypal lineages of Toxoplasma gondii', *Nucleic Acids Research*. doi: 10.1093/nar/gki604.

Khan, A. *et al.* (2006) 'Common inheritance of chromosome Ia associated with clonal expansion of Toxoplasma gondii', *Genome Research*. doi: 10.1101/gr.5318106.

Khan, A. *et al.* (2007) 'Recent transcontinental sweep of Toxoplasma gondii driven by a single monomorphic chromosome', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0702356104.

Khan, A., Miller, N., *et al.* (2011) 'A monomorphic haplotype of chromosome ia is associated with widespread success in clonal and nonclonal populations of Toxoplasma gondii', *mBio*, 2(6), pp. 1–10. doi: 10.1128/mBio.00228-11.

Khan, A., Dubey, J. P., *et al.* (2011) 'Genetic analyses of atypical Toxoplasma gondii strains reveal a fourth clonal lineage in North America', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2011.01.005.

Khan, A., Ajzenberg, D., *et al.* (2014) 'Geographic Separation of Domestic and Wild Strains of Toxoplasma gondii in French Guiana Correlates with a Monomorphic Version of Chromosome1a', *PLoS Neglected Tropical Diseases*. doi: 10.1371/journal.pntd.0003182.

Khan, A., Shaik, J. S., *et al.* (2014) 'NextGen sequencing reveals short double crossovers contribute disproportionately to genetic diversity in Toxoplasma gondii', *BMC Genomics*. doi: 10.1186/1471-2164-15-1168.

Khan, A. and Grigg, M. E. (2017) 'Toxoplasma gondii: Laboratory maintenance and growth', *Current Protocols in Microbiology*. doi: 10.1002/cpmc.26.

Khan, Asis *et al.* (2011) 'Genetic analyses of atypical Toxoplasma gondii strains reveal a fourth clonal lineage in North America', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2011.01.005.

Kim, K. and Weiss, L. M. (2004) 'Toxoplasma gondii: The model apicomplexan', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2003.12.009.

Kim, S.-K. and Boothroyd, J. C. (2005) 'Stage-Specific Expression of Surface Antigens by Toxoplasma gondii as a Mechanism to Facilitate Parasite Persistence', *The Journal of Immunology*. doi: 10.4049/jimmunol.174.12.8038.

Kim, S. K., Karasov, A. and Boothroyd, J. C. (2007) 'Bradyzoite-specific surface antigen SRS9 plays a role in maintaining Toxoplasma gondii persistence in the brain and in host control of parasite replication in the intestine', *Infection and Immunity*. doi: 10.1128/IAI.01862-06.

Kissinger, J. C. and DeBarry, J. (2011) 'Genome cartography: Charting the apicomplexan genome', *Trends in Parasitology*. doi: 10.1016/j.pt.2011.03.006.

Kitts, P. (2003) ‘Genome Assembly and Annotation Process’, in *The NCBI Handbook [Internet]*.

Koboldt, D. C. *et al.* (2010) ‘Challenges of sequencing human genomes’, *Briefings in Bioinformatics*. doi: 10.1093/bib/bbq016.

Kolben, M., Maurer, S. and Knitza, R. (2004) ‘Harninkontinenz bei einer frau’, *MMW-Fortschritte der Medizin*. Australian Society for Parasitology Inc., 146(29–30), pp. 59–60. doi: 10.1016/j.ijpara.2008.07.011.

Krishna, R. *et al.* (2015) ‘A large-scale proteogenomics study of apicomplexan pathogens-Toxoplasma gondii and Neospora caninum’, *Proteomics*, pp. 2618–2628. doi: 10.1002/pmic.201400553.

Kumar, S. (2005) ‘Molecular clocks: Four decades of evolution’, *Nature Reviews Genetics*. doi: 10.1038/nrg1659.

Kyes, S. A., Kraemer, S. M. and Smith, J. D. (2007) ‘Antigenic variation in Plasmodium falciparum: Gene organization and regulation of the var multigene family’, *Eukaryotic Cell*. doi: 10.1128/EC.00173-07.

Laliberté, J. and Carruthers, V. B. (2008) ‘Host cell manipulation by the human pathogen Toxoplasma gondii’, *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-008-7556-x.

Lau, Y. L. *et al.* (2016) ‘Deciphering the draft genome of Toxoplasma gondii RH strain’, *PLoS ONE*. doi: 10.1371/journal.pone.0157901.

Leckenby, A. and Hall, N. (2015) ‘Genomic changes during evolution of animal parasitism in eukaryotes’, *Current Opinion in Genetics and Development*. doi: 10.1016/j.gde.2015.11.001.

Lehmann, T. *et al.* (2000) ‘Strain typing of toxoplasma gondii:copmariosn of antigen-coding and housekeeping genes’, *Source Journal of Parasitology J. Parasitol*, 86(865), pp. 960–971. doi: 10.1645/0022-3395(2000)086[0960:STOTGC]2.0.CO;2.

Lehmann, T. *et al.* (2004) ‘Variation in the structure of Toxoplasma gondii and the roles of selfing, drift, and epistatic selection in maintaining linkage disequilibria’, *Infection, Genetics and Evolution*. doi: 10.1016/j.meegid.2004.01.007.

Lei, T. *et al.* (2014) ‘ROP18 is a key factor responsible for virulence difference between Toxoplasma gondii and Neospora caninum’, *PLoS ONE*. doi: 10.1371/journal.pone.0099744.

Lekutis, C. *et al.* (2001) ‘Surface antigens of Toxoplasma gondii: Variations on a theme’, *International Journal for Parasitology*. doi: 10.1016/S0020-7519(01)00261-2.

Li, L. *et al.* (2003) 'Gene discovery in the Apicomplexa as revealed by EST sequencing and assembly of a comparative gene database', *Genome Research*, 13(3), pp. 443–454. doi: 10.1101/gr.693203.

Lindsay, D. S. *et al.* (2006) 'Effects of High-Pressure Processing on Toxoplasma gondii Tissue Cysts in Ground Pork', *Journal of Parasitology*, 92(1), pp. 195–196. doi: 10.1645/GE-631R.1.

Lindsay, D. S. and Dubey, J. P. (1989) 'In vitro development of Neospora caninum (Protozoa: Apicomplexa) from dogs', *J Parasitol*. doi: 10.2307/3282960.

Lorenzi, H. *et al.* (2016) 'Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes', *Nature Communications*, 7. doi: 10.1038/ncomms10147.

Ma, L. *et al.* (2017) 'Neospora caninum ROP16 play an important role in the pathogenicity by phosphorylating host cell STAT3', *Veterinary Parasitology*. Elsevier, 243(2), pp. 135–147. doi: 10.1016/j.vetpar.2017.04.020.

Macêdo, A. G. *et al.* (2013) 'SAG2A protein from Toxoplasma gondii interacts with both innate and adaptive immune compartments of infected hosts', *Parasites and Vectors*. doi: 10.1186/1756-3305-6-163.

Medina-Esparza, L. *et al.* (2016) 'Genetic characterization of Neospora caninum from aborted bovine foetuses in Aguascalientes, Mexico', *Veterinary Parasitology*, 228, pp. 183–187. doi: 10.1016/j.vetpar.2016.09.009.

Meissner M., Reiss M., Viebig N., Carruthers VB., T. C., Tomavo S., A. J. and and Soldat D. (2002) 'A family of transmembrane microneme proteins of Toxoplasma gondii contain EGF-like domains and function as escorters', *Journal of Cell Science*. doi: 10.5194/isprsarchives-XXXIX-B7-45-2012.

de Melo Ferreira, A. *et al.* (2006) 'Genetic analysis of natural recombinant Brazilian Toxoplasma gondii strains by multilocus PCR-RFLP', *Infection, Genetics and Evolution*. doi: 10.1016/j.meegid.2004.12.004.

Mercier, C. and Cesbron-Delauw, M. F. (2015) 'Toxoplasma secretory granules: One population or more?', *Trends in Parasitology*. doi: 10.1016/j.pt.2014.12.002.

Michelin, A. *et al.* (2009) 'GRA12, a Toxoplasma dense granule protein associated with the intravacuolar membranous nanotubular network.Michelin, A., Bittame, A., Bordat, Y., Travier, L., Mercier, C., Dubremetz, J.-F., & Lebrun, M. (2009). GRA12, a Toxoplasma dense granule protein as', *International journal for parasitology*, 39(3), pp. 299–306. doi: 10.1016/j.ijpara.2008.07.011.

Milet, J. *et al.* (2010) 'Genome wide linkage study, using a 250K SNP map, of Plasmodium falciparum infection and mild malaria attack in a senegalese population', *PLoS ONE*. doi: 10.1371/journal.pone.0011616.

Minoche, A. E., Dohm, J. C. and Himmelbauer, H. (2011) 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems', *Genome Biology*. doi: 10.1186/gb-2011-12-11-r112.

Minot, S. *et al.* (2012) 'Admixture and recombination among Toxoplasma gondii lineages explain global genome diversity', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1117047109.

Montoya, J. G. and Liesenfeld, O. (2004) 'Toxoplasmosis', in *Lancet*, pp. 1965–1976. doi: 10.1016/S0140-6736(04)16412-X.

Morrison, D. A. (2009) 'Evolution of the Apicomplexa: where are we now?', *Trends in Parasitology*. doi: 10.1016/j.pt.2009.05.010.

Morrissette, N. S. and Sibley, L. D. (2002) 'Cytoskeleton of Apicomplexan Parasites', *Microbiology and Molecular Biology Reviews*. doi: 10.1128/MMBR.66.1.21-38.2002.

Mugnier, M. R., Stebbins, C. E. and Papavasiliou, F. N. (2016) 'Masters of Disguise: Antigenic Variation and the VSG Coat in Trypanosoma brucei', *PLoS Pathogens*, 12(9), pp. 1–6. doi: 10.1371/journal.ppat.1005784.

Nascimento, F. S. *et al.* (2016) 'Evaluation of library preparation methods for Illumina next generation sequencing of small amounts of DNA from foodborne parasites', *Journal of Microbiological Methods*. doi: 10.1016/j.mimet.2016.08.020.

Nath, A. and Sinai, A. P. (2003) 'Cerebral Toxoplasmosis', *Current Treatment Options in Neurology Current Science Inc*, 5, pp. 3–12.

Nevado, B., Ramos-Onsins, S. E. and Perez-Enciso, M. (2014) 'Resequencing studies of nonmodel organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics', *Molecular Ecology*, 23(7), pp. 1764–1779. doi: 10.1111/mec.12693.

Niedelman, W. *et al.* (2012) 'The rhoptry proteins ROP18 and ROP5 mediate Toxoplasma gondii evasion of the murine, but not the human, interferon-gamma response', *PLoS Pathogens*. doi: 10.1371/journal.ppat.1002784.

Niehus, S. *et al.* (2014) 'Virulent and avirulent strains of Toxoplasma gondii which differ in their glycosylphosphatidylinositol content induce similar biological functions in macrophages', *PLoS ONE*. doi: 10.1371/journal.pone.0085386.

Nowrousian, M. (2010) 'Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems', *Eukaryotic Cell*, 9(9), pp. 1300–1310. doi: 10.1128/EC.00123-10.

Okamoto, N. and McFadden, G. I. (2008) 'The mother of all parasites', *Future Microbiology*. doi: 10.2217/17460913.3.4.391.

Ong, Y. C., Reese, M. L. and Boothroyd, J. C. (2010) 'Toxoplasma Rhoptry Protein 16 (ROP16) subverts host function by direct tyrosine phosphorylation of STAT6', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M110.112359.

Paibomesai, M. I. *et al.* (2010) 'Clock genes and their genomic distributions in three species of salmonid fishes: Associations with genes regulating sexual maturation and cell cycling', *BMC Research Notes*, 3, pp. 1–21. doi: 10.1186/1756-0500-3-215.

Pain, A. *et al.* (2005) 'Genome of the host-cell transforming parasite Theileria annulata compared with T. parva', *Science*. doi: 10.1126/science.1110418.

Pain, A. *et al.* (2008) 'The genome of the simian and human malaria parasite Plasmodium knowlesi', *Nature*. doi: 10.1038/nature07306.

Panunzi, L. G. and Agüero, F. (2014) 'A Genome-Wide Analysis of Genetic Diversity in Trypanosoma cruzi Intergenic Regions', *PLoS Neglected Tropical Diseases*. doi: 10.1371/journal.pntd.0002839.

Pappas, G., Roussos, N. and Falagas, M. E. (2009) 'Toxoplasmosis snapshots: Global status of Toxoplasma gondii seroprevalence and implications for pregnancy and congenital toxoplasmosis', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2009.04.003.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, Andrew J, *et al.* (2010) 'Antagonistic coevolution accelerates molecular evolution', *Nature*. Nature Publishing Group, 464(7286), pp. 275–278. doi: 10.1038/nature08798.

Peng, X. *et al.* (2015) 'Re-alignment of the unmapped reads with base quality score', *BMC Bioinformatics*. doi: 10.1186/1471-2105-16-S5-S8.

Petter, M. and Duffy, M. F. (2015) 'Antigenic variation in plasmodium falciparum', in *Results and Problems in Cell Differentiation*. doi: 10.1007/978-3-319-20819-0_3.

Phillippy, A. M. (2017) 'New advances in sequence assembly', *Genome Research*. doi: 10.1101/gr.223057.117.

Pollard, A. M. *et al.* (2008) 'Highly polymorphic family of glycosylphosphatidylinositol-anchored surface antigens with evidence of developmental regulation in Toxoplasma gondii', *Infection and Immunity*. doi: 10.1128/IAI.01170-07.

Radke, J. R. *et al.* (2001) 'Defining the cell cycle for the tachyzoite stage of Toxoplasma gondii', *Molecular and Biochemical Parasitology*. doi: 10.1016/S0166-6851(01)00284-5.

Ramakrishnan, C. *et al.* (2017) 'The merozoite-specific protein, TgGRA11B, identified as a component of the Toxoplasma gondii parasitophorous vacuole in a tachyzoite expression model', *International Journal for Parasitology*, 47(10–11), pp. 597–600. doi: 10.1016/j.ijpara.2017.04.001.

Ramaprasad, A. *et al.* (2015) 'Comprehensive evaluation of Toxoplasma gondii VEG and Neospora caninum LIV genomes with tachyzoite stage transcriptome and proteome defines novel transcript features', *PLoS ONE*. doi: 10.1371/journal.pone.0124473.

Raz, Y. and Tannenbaum, E. (2010) 'The influence of horizontal gene transfer on the mean fitness of unicellular populations in static environments', *Genetics*. doi: 10.1534/genetics.109.113613.

Raza, S., Shoaib, M. W. and Mubeen, H. (2016) 'Genetic Markers: Importance, uses and applications', *International Journal of Scientific and Research Publications*.

Regidor-Cerrillo, J. *et al.* (2006) 'Multilocus microsatellite analysis reveals extensive genetic diversity in Neospora caninum.', *The Journal of parasitology*, 92(3), pp. 517–524. doi: 10.1645/GE-713R.1.

Regidor-Cerrillo, J. *et al.* (2013) 'Genetic Diversity and Geographic Population Structure of Bovine Neospora caninum Determined by Microsatellite Genotyping Analysis', *PLoS ONE*, 8(8). doi: 10.1371/journal.pone.0072678.

Reid, A. J. *et al.* (2012) 'Comparative genomics of the apicomplexan parasites Toxoplasma gondii and neospora caninum: Coccidia differing in host range and transmission strategy', *PLoS Pathogens*. doi: 10.1371/journal.ppat.1002567.

Reid, A. J. (2015) 'Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa.', *Parasitology*, (Suppl 1). doi: 10.1017/S0031182014001528.

Ricklefs, R. E. and Outlaw, D. C. (2010) 'A molecular clock for malaria parasites', *Science*. doi: 10.1126/science.1188954.

Risco-Castillo, V. *et al.* (2011) 'Identification of a gene cluster for cell-surface genes of the SRS superfamily in Neospora caninum and characterization of the novel SRS9 gene', *Parasitology*, 138(14), pp. 1832–1842. doi: 10.1017/S0031182011001351.

Rizzo, J. M. and Buck, M. J. (2012) 'Key principles and clinical applications of "next-generation" DNA sequencing', *Cancer Prevention Research*. doi: 10.1158/1940-6207.CAPR-11-0432.

Robert-Gangneux, F. and Dardé, M. L. (2012) 'Epidemiology of and diagnostic strategies for toxoplasmosis', *Clinical Microbiology Reviews*, pp. 264–296. doi: 10.1128/CMR.05013-11.

Saeij, J. P. J., Boyle, J. P. and Boothroyd, J. C. (2005) 'Differences among the three major strains of Toxoplasma gondii and their specific interactions with the infected host', *Trends in Parasitology*. doi: 10.1016/j.pt.2005.08.001.

Salehi, N., Gottstein, B. and Haddadzadeh, H. R. (2015) 'Genetic diversity of bovine Neospora caninum determined by microsatellite markers', *Parasitology International*, 64(5). doi: 10.1016/j.parint.2015.05.005.

Sandmann, S. *et al.* (2017) 'Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data', *Scientific Reports*. doi: 10.1038/srep43169.

Schares, G. *et al.* (1998) 'The efficiency of vertical transmission of Neospora caninum in dairy cattle analysed by serological techniques', *Veterinary Parasitology*, 80(2), pp. 87–98. doi: 10.1016/S0304-4017(98)00195-2.

Schock, A. *et al.* (2001) 'Genetic and biological diversity among isolates of Neospora caninum', *Parasitology*, 123(1), pp. 13–23. doi: 10.1017/S003118200100796X.

Schuster, S. C. (2008) 'Next-generation sequencing transforms today's biology', *Nature Methods*. doi: 10.1038/nmeth1156.

Seeber, F. and Steinfelder, S. (2016) 'Recent advances in understanding apicomplexan parasites', *F1000Research*. doi: 10.12688/f1000research.7924.1.

Sharif, M. *et al.* (2017) 'Genetic diversity of Toxoplasma gondii isolates from ruminants: A systematic review', *International Journal of Food Microbiology*. doi: 10.1016/j.ijfoodmicro.2017.07.007.

Shwab, E. K. *et al.* (2014) 'Geographical patterns of Toxoplasma gondii genetic diversity revealed by multilocus PCR-RFLP genotyping', *Parasitology*. doi: 10.1017/S0031182013001844.

Shwab, E. K. *et al.* (2016) 'The ROP18 and ROP5 gene allele types are highly predictive of virulence in mice across globally distributed strains of Toxoplasma gondii', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2015.10.005.

Sibley, L. D. *et al.* (2002) 'Genetic approaches to studying virulence and pathogenesis in Toxoplasma gondii', in *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi: 10.1098/rstb.2001.1017.

Sibley, L. David *et al.* (2009) 'Forward genetics in Toxoplasma gondii reveals a family of rhoptry kinases that mediates pathogenesis', *Eukaryotic Cell*, 8(8), pp. 1085–1093. doi: 10.1128/EC.00107-09.

Sibley, L David *et al.* (2009) 'Genetic diversity of Toxoplasma gondii in animals and humans.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1530), pp. 2749–61. doi: 10.1098/rstb.2009.0087.

Sibley, L. D. and Boothroyd, J. C. (1992) 'Virulent strains of Toxoplasma gondii comprise a single clonal lineage.', *Nature*, 359(6390), pp. 82–5. doi: 10.1038/359082a0.

Sidik, S. M. *et al.* (2016) 'A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes', *Cell*, 166(6). doi: 10.1016/j.cell.2016.08.019.

Sidik, S. M., Huet, D. and Lourido, S. (2018) 'CRISPR-Cas9-based genome-wide screening of toxoplasma gondii', *Nature Protocols*. doi: 10.1038/nprot.2017.131.

Sinai, A. P. and Joiner, K. A. (2001) 'The Toxoplasma gondii protein ROP2 mediates host organelle association with the parasitophorous vacuole membrane', *Journal of Cell Biology*. doi: 10.1083/jcb.200101073.

Singh, R. R. (2017) 'Next generation sequencing technologies', in *Comprehensive Medicinal Chemistry III*. doi: 10.1016/B978-0-12-409547-2.12327-3.

Su, C. *et al.* (2003) 'Recent expansion of Toxoplasma through enhanced oral transmission', *Science*. doi: 10.1126/science.1078035.

Su, C. *et al.* (2004) 'Typing single-nucleotide polymorphisms in Toxoplasma gondii by allele-specific primer extension and microarray detection.', *Methods Mol Biol*. doi: 10.1021/ac071008v.

Su, C. *et al.* (2012) 'Globally diverse Toxoplasma gondii isolates comprise six major clades originating from a small number of distinct ancestral lineages', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1203190109.

Su, C., Zhang, X. and Dubey, J. P. (2006) 'Genotyping of Toxoplasma gondii by multilocus PCR-RFLP markers: A high resolution and simple method for identification of parasites', *International Journal for Parasitology*. doi: 10.1016/j.ijpara.2006.03.003.

Sullivan, W. J. and Jeffers, V. (2012) 'Mechanisms of Toxoplasma gondii persistence and latency', *FEMS Microbiology Reviews*. doi: 10.1111/j.1574-6976.2011.00305.x.

Takemae, H. *et al.* (2013) 'Characterization of the interaction between Toxoplasma gondii rhoptry neck protein 4 and host cellular β-tubulin', *Scientific Reports*, 3, pp. 1–9. doi: 10.1038/srep03199.

Talevich, E. and Kannan, N. (2013) 'Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors.', *BMC evolutionary biology*, 13(1). doi: 10.1186/1471-2148-13-117.

Taylor, S. *et al.* (2006) 'A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen Toxoplasma gondii', *Science*. doi: 10.1126/science.1133643.

Tenter, A. M., Heckeroth, A. R. and Weiss, L. M. (2000) 'Toxoplasma gondii: from animals to humans.', *International journal for parasitology*, 30(12–13), pp. 1217–58. doi: 10.1016/S0020-7519(00)00124-7.

Thankaswamy-Kosalai, S., Sen, P. and Nookaew, I. (2017) 'Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics', *Genomics*. doi: 10.1016/j.ygeno.2017.03.001.

Tonkin, M. L. *et al.* (2010) 'Structure of the micronemal protein 2 A/I domain from Toxoplasma gondii', *Protein Science*, 19(10), pp. 1985–1990. doi: 10.1002/pro.477.

Treangen, T. J. and Salzberg, S. L. (2013) 'Repetitive DNA and next-generation sequencing: computational challenges and solutions', *Nat Rev Genet.*, 13(1), pp. 36–46. doi: 10.1038/nrg3117.Repetitive.

Tyler, J. S., Treeck, M. and Boothroyd, J. C. (2011) 'Focus on the ringleader: The role of AMA1 in apicomplexan invasion and replication', *Trends in Parasitology*. doi: 10.1016/j.pt.2011.04.002.

Usman, T. *et al.* (2017) 'Unmapped reads from cattle RNAseq data: A source for missing and misassembled sequences in the reference assemblies and for detection of pathogens in the host', *Genomics*. doi: 10.1016/j.ygeno.2016.11.009.

Vignal, A. *et al.* (2002) 'A review on SNP and other types of molecular markers and their use in animal genetics', *Genetics Selection Evolution*. doi: 10.1186/1297-9686-34-3-275.

Walzer, K. A. *et al.* (2014) 'Hammondia hammondi harbors functional orthologs of the host-modulating effectors GRA15 and ROP16 but is distinguished from toxoplasma gondii by a unique transcriptional profile', *Eukaryotic Cell*, 13(12). doi: 10.1128/EC.00215-14.

Walzer, K. A. and Boyle, J. P. (2012) 'A single chromosome unexpectedly links highly divergent isolates of Toxoplasma gondii', *mBio*. doi: 10.1128/mBio.00284-11.

Wang, J. L. *et al.* (2017) 'Functional characterization of rhoptry kinome in the virulent Toxoplasma gondii RH strain', *Frontiers in Microbiology*, 8(JAN). doi: 10.3389/fmicb.2017.00084.

Wang, P. Y. *et al.* (2012) 'Genetic diversity among Toxoplasma gondii isolates from different hosts and geographical locations revealed by analysis of ROP13 gene sequences', *African Journal of Biotechnology*, 11(25), pp. 6662–6665. doi: 10.5897/AJB12.056.

Wasmuth, J. *et al.* (2009) 'The origins of apicomplexan sequence innovation', *Genome Research*. doi: 10.1101/gr.083386.108.

Wasmuth, J. D. *et al.* (2012) 'Integrated bioinformatic and targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of toxoplasma virulence', *mBio*. doi: 10.1128/mBio.00321-12.

Van Der Weide, R. H. *et al.* (2016) 'The genomic scrapheap challenge; extracting relevant data from unmapped whole genome sequencing reads, including strain specific genomic segments, in rats', *PLoS ONE*. doi: 10.1371/journal.pone.0160036.

Wellems, T. E., Hayton, K. and Fairhurst, R. M. (2009) 'The impact of malaria parasitism: From corpuscles to communities', *Journal of Clinical Investigation*. doi: 10.1172/JCI38307.

Whitacre, L. K. *et al.* (2015) 'What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual', *BMC Genomics*, 16(1). doi: 10.1186/s12864-015-2313-7.

Xiao, F. *et al.* (2019) 'An accurate and powerful method for copy number variation detection', *Bioinformatics*. doi: 10.1093/bioinformatics/bty1041.

Yu, X. and Sun, S. (2013) 'Comparing a few SNP calling algorithms using low-coverage sequencing data', *BMC Bioinformatics*. doi: 10.1186/1471-2105-14-274.

Zhou, J. *et al.* (2011) 'Copy-number variation: The balance between gene dosage and expression in Drosophila melanogaster', *Genome Biology and Evolution*. doi: 10.1093/gbe/evr023.

# Appendices

All the appendices include in this thesis are available in the following website **http://datacat.liverpool.ac.uk/id/eprint/743**

- **The dataset includes:**
  - **Dataset of chapter three (A)** include**:** summary of the OrthoMCL clustering of *T. gondii* and *N. caninum.*
  - **Dataset of chapter four (B)** include; B1(*NC-Bahia*), B2(*NC-1*) and B3(*NC-Liverpool)* that have a high impacts of unique SNPs per strain.
  - **Dataset of chapter five (C)** include; C1(*T. GT1*), C2(*T. MAS*), C3 (*T. P89*), C4(*T. CAST*), C5(*T. VEG*) and C6 (*T. COUG*) that have a high impacts of unique SNPs per strain.